Guided by Objective Metrics:

reducing researcher bias in biomedical image analysis

Author:

Luke William TERNES

Supervisor:

Dr. Young Hwan CHANG

A DISSERTATION

Presented to the

Department of Biomedical Engineering

Oregon Health & Science University

School of Medicine

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

March 15, 2022

Contents

List of Figures

List of Tables

List of Abbreviations

2D	**2**-**D**imensional
3D	**3**-**D**imensional
ADM	**A**cinar-**T**o-**D**uctal **M**etaplasia
AMY	**A**mylase
BCE	**B**inary **C**ross **E**ntropy
CCA	**C**anonical **C**orrelation **A**nalysis
CRC	**C**olo**R**ectal **C**ancer
C-VAE	**C**onditional **V**ariational **A**uto**E**ncoder
CODEX	**CO**-**DE**tection by inde**X**ing
XAE	**Cross**-domain **A**uto**E**ncoder
CyCIF	**Cy**Clic **I**mmuno**F**luorescence
DDR	**D**NA **D**amage and **R**epair
ES	**E**ffect **S**ize
ELBO	**E**vidence **L**ower **BO**und
EGFR	**E**pidermal **G**rowth **F**actor **R**eceptor
EMT	**E**pithelial-to-**M**esenchymal **T**ransition
GAN	**G**enerative **A**dversarial **N**etwork
H&E	**H**ematoxylin and **E**osin

IF	**I**mmuno**F**luorescence
JSD	**J**ensen-**S**hannon **D**ivergence
KL	**K**ullback-**L**eibler
MSE	**M**ean **S**quared **E**rror
MxIF	**M**ultiple**x** **I**mmuno**F**luorescence
MTI	**M**ultiplex **T**issue **I**maging
ME-VAE	**M**ulti-**E**ncoder **V**ariational **A**uto**E**ncoder
NMI	**N**ormalized **M**utual **I**nformation
PDAC	**P**ancreatic **D**uctal **A**deno**C**arcinoma
PanIN	**Pan**creatic **I**ntraepithelial **N**eoplasia
PanK	**Pan**-**K**eratin
ROC	**R**eceiver **O**perating **C**haracteristic
ROI	**R**egion **O**f **I**nterest
RPPA	**R**everse **P**hase **P**rotein **A**ssay
SHIFT	**S**peedy **H**istological-to-**I**mmuno**F**luorescent **T**ranslation
SSIM	**S**tructural **SIM**ilarity
TMA	**T**issue **M**icro**A**rray
tp	**T**ukey-pairwise **p**-value
VAE	**V**ariational **A**uto**E**ncoder
VISTA	**VI**sual **S**emantic **T**issue **A**nalysis
WSI	**W**hole **S**lide **I**mage

Abstract

An increasing amount of attention is being given to the structural and morphological aspects of biological function, and by extension the computational challenges necessary to understand them are becoming increasingly important. In cancer research, structural changes in the tumor microenvironment and state changes in the cells, as well as the spatial distributions of proteins at both the tissue and cellular levels are key to understanding cancer progression and evolution. Imaging modalities are integral to the development of morpho-spatial analysis because they allow us to capture the structural information in tangent to the expression information across many different biological scales with high resolution. Multiplex imaging specifically has allowed researchers to capture an incredible amount of data for dozens of proteins at a time, but the computational methods that are necessary to comprehend the true diversity of the information that is contained in the data are only now being developed. Digital pathology applications suffer from a lack of consistent, reproducible, and unbiased methods, and as a result the interpretations of the results are prone to inter-operator and inter-institution variability. These limiting methods include: 1) imperfect manual annotation and segmentation which are both slow and produce different results for every operator, 2) the reliance upon classical morpho-spatial features which are generalist and fail to quantify novel and complex features of biological importance, 3) the subjective selection of regions-of-interest and multiplex marker panels, both of which will vary between researchers and cannot accurately be done without substantial prior information. Because deep learning can capture complex information from large data and the results of such models are reproducible regardless of operator, they

provide the opportunity to address many of the limitations that multiplex imaging faces. First, I propose a deep learning model for virtual semantic segmentation of nuanced features (VISTA) as a solution to slow and imperfect annotation in pancreatic ductal adenocarcinoma tissue samples and discuss how similar pipelines can be developed for use in new pathologies. Second, I propose a novel multi-encoder variational autoencoder (ME-VAE) architecture, which is capable of extracting biologically relevant morpho-spatial features from single cell images, specific to each dataset and without the bias of traditional imaging features. Finally, I propose a series of deep learning methods developed for reducing researcher burden and bias in multiplex imaging by reconstructing 3-dimensional tissue volumes, selecting representative regions-of-interest using convex optimization, and decreasing panel sizes by calculating a theoretically ideal reduced panel capable of imputing all the information in the original full panel. Implementation of these methods will help to advance the growing community of artificial intelligence research in the biomedical domain and allow researchers to reproducibly quantify the morpho-spatial data held in their images, which will lead to novel breakthroughs in the cancer domain and beyond.

Acknowledgements

My success over the past several years would not have been possible without the continued support and encouragement of my mentor, Dr. Young Hwan Chang, whose excessive patience and tempered guidance have helped me overcome many of the challenges of life in research. It was his hard work ethic that I tried my best to reproduce in my own life as I developed my career. I would not be where I am today were it not for him, and I am truly grateful.

I owe an unquantifiable amount of appreciation to the members of my lab, Dr. Erik Burlingame, Dr. Geoffrey Schau, and Elliot Gray, who gave me the traction I needed when I was first starting out. They taught me more about deep learning than I could have ever learned from reading literature. Thanks for being my closest friends over the years.

I would also like to thank all my collaborators and the members of my advisory committees, Dr. Laura Heiser, Dr. Gordon Mills, Dr. Joe W. Gray, Dr. Daniel Zuckerman, Dr. Guillaume Thibault, and Dr. John Muschler, for their wise counseling despite my incessant pestering over the years. My thesis likely would never have come together without the instruction and opportunities they gave me.

This is dedicated to whoever finally cures cancer and makes it so I don't have to keep doing this anymore. Thank you.

Chapter 1

Introduction

Great knowledge sees all in one.
Small knowledge breaks down into the many.

Readings from Chuang Tzu

1.1 The use of biomedical imaging

Images are one of the most rapid and efficient ways to represent and convey intelligible information, and as such the ability to visualize disease states is essential in interpreting the biology. Visualization of biology is not always possible with the naked eye either because it is too small, because it is out of reach, or because the biology of interest is not visible without specific treatment. Biomedical imaging helps to address this by allowing physicians and researchers to extract dense, biologically relevant information that would normally be inaccessible through the use of tissue sections, advanced imaging devices, and/or biologically target staining to highlight certain features. Imaging is a well-established practice in the medical

field and has been a staple of medical diagnostics for hundreds of years. Hematoxylin and eosin (H&E) staining in particular has been a standard imaging practice since 1876[150] and has allowed physicians to make decisions regarding diagnosis, grading, and prognosis ever since. Despite its age, H&E imaging continues to be consistently used, even in the presence of other more complex imaging modalities because it is cheap and fast. It only uses two stains but it allows physicians and researchers to interpret high-level features of biology such as tissue organization, type, and morphology.

This ability to examine and measure the morphological and spatial (morphospatial) components of disease is why imaging is so important. While many modern tests and modalities, such as those in the realm of -omics (RNA sequencing[92], reverse-phase protein arrays[50], chromatin sequencing[14]), can provide large amounts of expression data, they currently can do little to tell us about the structural and organizational patterns of cells within a tissue as well as the distribution of the expressions within cells. Imaging is unique in that it preserves the spatial by aspect of information and therefore spatially resolve the similar expression profiles to the other modalities. In cancer, this spatial aspect of information and expression has been shown to be an important step in cancer subtyping and treatment[26]. For example, recent discoveries have found that the proximities and spatial distributions of many cell types such as fibroblasts, immune cells, and tumor cells have an influence on clinical outcomes and patient survival [81, 10].

The importance of imaging in the medical field has pushed researchers to develop

more and more powerful tools that can obtain images with higher spatial resolution and increased profiling capacities, capable of now staining for dozens of stains on the same tissue and cells. These are broadly called state-of-the-art multiplex imaging modalities and include such methods as cyclic immunofluorescence (CyCIF)[64], multiplex ion beam imaging (MIBI)[4], multiplex immunohistochemistry (mIHC)[131], and co-detection by indexing (CODEX)[38]. Each modality, understandably, comes with its own benefits and drawbacks. Fluorescence-base methods are able to image large areas of tissue but suffer from auto-fluorescence issues as a result of fixation[64, 38, 36], while mass spectrometry-based approaches are able to achieve higher signal-to-background ratios while only being able to image smaller areas[4, 37]. Additionally enzyme-based antibody methods like mIHC[131] have more harsh label stripping conditions which results in increased deleterious tissue effects. These multiplex modalities have increased the amount of data being gathered both in terms of the spatial resolution (some being able to image whole slide images at the single cell level) and marker diversity (some being able to image the same tissue with up to 100 stains). The problem is that researchers these modalities produces increasingly large amounts of data with more complexities, interactions, and confounding factors. The best methods to analyze these imaging features, however, are still the topic of research and testing, and as of yet there is no current standard for many of the challenges.

1.2 Limitations of the imaging modalities

Although multiplexed imaging technologies like CyCIF[64] have introduced many benefits to researchers, they are still relatively novel tools that have not yet been fully optimized, and as such they each come with their own downsides that restrict their broad and rapid application. Firstly, highly multiplexed technologies are slow and labor intensive technologies[64, 38, 36, 4, 37]. CyCIF, for example, is a process involving many iterative rounds of staining, imaging, and quenching and can take weeks to complete depending on the size of the tissue and the number of stains being used[64]. Some technologies, like multiplex immunohistochemistry[131], can operate in a matter of days, but even this is too slow for broad deployment in settings where pathologists and doctors need rapid results to dictate immediate treatment. The faster multiplex platforms, however, typically come with decreased resolution, multiplex stain capacity, or throughput[131, 4, 37]. These time estimates also only take into consideration the amount of time required to actually capture the image following all the upstream preparation of the sample and panel design, the processes of which each come with limitations of their own.

Panel design is a subjective process that is heavily reliant on prior literature, researcher experience, and many iterations of experimentation to optimize with no guarantee that the panel will capture all the relevant biology[94, 32]. The process requires balancing the selection of specific biological targets with the practical limitations of the physical biology: stains may or may not compete within a single round of staining or may have off-target effects[135, 6]. When designing the panel, researchers must use a substantial amount of prior knowledge to determine which

stains will produce the most amount of biologically relevant data, but it can be difficult to predict which stains will be interesting beforehand and whether specific stains even need to be included because they share mutual information with the other stains. As a result, multiple researchers are able to select reasonable and distinct panel sets for an identical experiment, with no quantitative metric of which will be better beforehand. Although the multiplex technologies enable staining with significantly more depth, conducting many subsequent imaging rounds can result in increased autofluorescence within the image (as a result of interactions when blocking with normal serum) and increased degradation of the tissue[53]. For this reason, researchers must be careful to only select markers that will be biologically important because although the panel size has been increased, the space on the panel is still limited. The fact that these decisions must be made with each experiment and dataset restricts the application of multiplex imaging away from domains where 1) there may be a lack of prior information and 2) testing and deployment must be rapid.

Even outside of multiplex imaging, the medical field is slowed by image-based computational processes. It is true that imaging data comes with unique and important morpho-spatial information, but unlike other expression measurements, these morpho-spatial patterns do not come already labeled and quantified. Many imaging features are incredibly complex and cannot be captured with standard handcrafted feature metrics for stain expression and morphology, as is the case for subtle histological features of cancer progression[115] (further discussed in chapter 2) and there is bias when selecting features, the result of which can ignore biologically relevant features and put inappropriate focus on other technical features

(further discussed in chapter 3). Currently, many tasks require researchers to manually process images, by annotating tumors, labeling cell types of interest, and quantifying/normalizing stains, all of which require a significant amount of time and are subjective decisions prone to inter- and intra-operator variation[31, 149, 27, 111, 59]. Sometimes, there aren't even appropriate methods for bulk annotation, and cell type labeling is often performed simply using clustering, whereby cell data is computationally grouped based on similarity, or binarized marker gating, whereby cells are called as being positive or negative for protein expression using a simple threshold.

These processes of annotation are tedious and waste the time that researchers and physicians. The issue of annotation is compounded when one looks at the size of datasets, which are only increasing in volume as we improve the technology. Whole slide images now contain millions of cells for labeling[67], multiplex data uses dozens of stains that each independently require thresholding[64], and there is an increasing demand to research 3-dimensional tissue volumes[60, 70, 67]. Not only do these trends increase manual workload, but whole process must be repeated for every new dataset that is being analyzed. Cost is also a limitation of these new platforms, making their use prohibitively expensive for small or poorly funded labs. Additionally, the cost of researchers, experts, computationalists, and supplemental resources will also increase as time is spent on menial tasks that do not necessarily require innovative thought.

1.3 The limitations of imaging by eye

While the technology that is used for imaging will most certainly continue to improve over the coming years and reduce the aforementioned limitations, one aspect that will continue to be a limitation on image analysis is the human element. Deep learning models are capable of parallel processing, allowing a single model to analyze multiple images simultaneously, completing tasks in a fraction of the time it would take their human counterparts[126]. Moreover, once a model has been successfully trained, the computational resources necessary are generally fairly small, capable of being deployed on generic smart phones, as can be seen in the many facial recognition tools used in today's social media applications. For this reason, deep learning image models have and will continue to be deployed en masse throughout the industrial setting and will grow more and more prevalent within the medical field in the coming years.

An added benefit that computational algorithms have over the human analyst is consistency. Not only has analysis by multiple experts been shown to be vulnerable to inter-observer variation and bias [31, 149, 27], but even a single expert will produce different results between multiple viewings of an image, even for simple tasks such as counting cells[111, 59]. Because computation is deterministic, the predictions from a deep learning model will produce the same results every time. Although the computational algorithms may still be prone to error, the errors are reproducible, diagnosable, and with some effort can be fixed/improved in subsequent iterations. The errors made by human annotation will often go unaddressed

with no good way to overcome limitations in an inherently biased and unchangeable biological neural network.

When a human views an image by eye, they can understand the whole of the image, but when they are given the same image data in the form of intensity values and matrices, it becomes completely unintelligible. Additionally, current visualization methods struggle to convey multiplex images since computer visualization is limited to 3 primary color channels (red,green,blue)[116] and multiplex images can have dozens of channels to visualize simultaneously[64]. Extracting hidden patterns from big data requires the ability to parse through more information than can be held in the human attention at one time, which is something machines are very good at doing, and as a result recent advanced computational methods have been shown to extract hidden patterns from images that are imperceptible to humans in the biological domain[78, 133].

Similarly, machine learning allows researchers to apply quantitative metrics to features that the human brain can only do qualitatively. Using features in cancer images, pathologists can make qualitative decisions in order to classify tumor grade, but these assignments do not reflect the continuous nature of cancer progression, would be better captured in continuous quantitative values, and lack mathematical certainty. When researching topics which require nuance in tissue and cell state, such as developmental mechanisms of disease progression, these qualitative decisions become even more perplexing, as there is often no clear divide between one state and another[77]. Cell segmentation is commonly done with semantic segmentation, performed by labeling positive area at the pixel level or instance segmentation, performed by selecting regions containing a single cell. Even the

more simple semantic segmentation-based tasks, however, suffer from imprecise ground truth as it can be difficult to determine where one cell begins and another ends, especially when multiple cells overlap. As a result, calls and decisions made between researchers will not always be consistent[31, 149, 27, 111, 59]. Many basic tissue level segmentation and quantification methods for the purpose of separating these disease states and tissue types rely on the thresholding of stains to make binary calls[123, 69]. These thresholding methods, however, are prone to many sources of variation and bias. The manually determined thresholds will vary by user, and the tissue images often have uneven staining and fluorescence, which further confounds the process[126].

Many morpho-spatial features that are important for biomedical analysis, for example the texture of a surface, are things that humans can qualitatively observe, but lack the capacity to empirically quantify. Classical feature sets[138, 80], which are defined by static handcrafted metrics, can extract some morpho-spatial features but these lack the complexity to extract all the rich information from multiplex images which can be defined by n-dimensional pixel level interactions. To further add to the complexity of analysis, tissue level multiplex images can be incredibly large, comprised of hundreds of billions of pixels and millions of cells[67]. This necessitates the trimming down of data into more comprehensible regions, as oftentimes many regions of the image might not be of biological interest[67]. This could be because some regions lack pathologically relevant tissue, do not contain rare/novel cell types, or are not undergoing the specific micro-environmental changes examined in the study. Both of these processes require a substantial amount of prior information. Moreover, the decisions do not have empirical backing to show that

they optimize the amount of relevant information gained from the experiment, and the choices for each part are subject to researcher bias. Empirical methods for selection important regions have been performed in H&E[95], but these attempts have been very limited, capturing only obvious and high-level tissue features which would not prove useful for more nuanced features at the cellular level, which is necessary for multiplex imaging modalities.

1.4 Applications and limitations of existing machine learning methods in biomedical imaging

Machine learning is a subset of artificial intelligence methods that enable the rapid and accurate analysis of big data without having to be specifically programmed for the task[28]. Using provided input data, the machine learning methods learn to complete a specific task by "programming" themselves in a process called "training". In doing so, the models learn a specific set of parameters, weights, and features, that optimize that performance of their specific task. In order to complete the same tasks normally, researchers would design similar parameters a weights in a sub-optimal process of trail and error.

Depending on how the data is given to the model and how the process of training is conducted, machine learning can be classified into different types. Although there are many distinctions, supervised learning and unsupervised learning are of particular relevance to this work. Supervised learning is when the machine learning model is trained using classified, labeled, or ground truth data, such that the model

the model has an intended output that it can compare its predictions to. Unsupervised learning is when the model does not receive labeled data during training, and therefore must infer information regarding the data that is not given to the model directly[28]. Additionally, deep learning is an subset of machine learning models that is characterized by having multiple layers in its architecture, making it capable of performing more complex tasks while requiring more computational resources to train and use[47].

Hundreds of new deep learning architectures are being developed and published every year, with more than 700 peer-reviewed AI publications being produced within the US for the medical field alone in 2019 according to Stanford University's AI Index Report[154]. From among these, there are several common deep learning templates that have remained relevant without drowning in the quickly changing sea of "state-of-the-art". Although addressing all of them is outside the scope of this work, I will briefly describe three that are highly relevant to the projects detailed later.

Segmentation is a vital part of the image analysis pipeline for most cancer research; specific tissues and cells can only be analyzed individually if they are first segmented from the rest of the image. The UNet[102]is an architecture for neural networks using only convolutional layers to compress and expand the image in such a way that it can generate the desired output, commonly segmentation masks. The UNet has served as the baseline template for deep learning segmentation, and over the years has seen many updates and modifications to improve its performance[19, 151, 52]. One of the key advances in this domain has been the creation of widely applicable cell segmentation methods such as Cellpose[120], Mesmer[40],

and Stardist[147], which can perform a variety of cell segmentation tasks with little expertise required. These generalist methods allow for the deployment of a single reproducible deep learning model on many different datasets, which can be characterized by different cell types, panels, and labs of imaging, without the need for researchers to train their own models or design the model specific to their data. Although these show promise and widespread use, the application of generalist models is restricted to single cell and nuclear segmentations, where there is a common and easily defined objective shared by most researchers. As of yet, there are no successful and widespread generalist methods for tissue level features, and currently tissue level segmentation requires manual annotation by researchers or the expertise to train highly specific models.

Once the targets of interest have been segmented from an image, it is necessary to extract biologically relevant features that can help researchers understand what is happening to the cell or tissue. Although many common handcrafted features and metrics exist for this purpose[138, 80], there are many more complicated features that still need to be extracted and are not readily perceivable by the human eye[78]. One tool commonly used to extract such features without researcher bias is the Variational Autoencoder (VAE)[58], which compresses images into a series of quantitative values that describe all the features within the image and then reconstructs the original image from said vector in order to enforce that the representative values are relevant. The logic behind the model is that if you can adequately reconstruct the image using the encoded values, then the values must contain all the necessary the information about the image. Using a VAE not only allows researchers to capture novel information not adequately represented in handcrafted

metrics, but does so in an undirected and unbiased fashion[58, 78]. This means that researchers can extract relevant information without having to know the features of interest beforehand.

VAEs, however, come with their own limitations. This includes interpretability. Because many features are quantified in an overlapping and interconnected fashion, it can be difficult to disentangle the biological meanings that come out of a black box[44, 86]. VAEs are also reliant on the fine tuning of many parameters, and small changes in these parameters can completely change the way the models learn or even cause the model to collapse, meaning that the model converges to state where the encodings and predictions are meaningless in relation to the input but optimize the loss function nonetheless[44]. Relevant to the work described here, VAEs can be hypersensitive to transformation features, such as rotation, skew, and scale, which are descriptors of the image but depending on the context might contain no relevant biology, and this hypersensitivity inhibits the ability of models to fully learn biological representations[35, 48, 8, 155, 84, 42]. This has been a topic of considerable research, but most methods designed to overcome this hypersensitivity are only directed at single features of disinterest at a time [48, 8, 155, 84, 42, 106, 93].

Using the information and features available from imaging data, researchers often require the ability to predict realistic data that they do not currently have. Generative Adversarial Networks (GANs) allow for generator networks (like UNets) to be pushed toward producing more realistic results necessary for synthetic data generation. It does this by coupling a discriminator network to the generator, which punishes the generator if it produces results that can be distinguished from real

data. Although commonly applied for synthetic image prediction[114, 33, 133], their adversarial concept has also been applied to segmentation[140] and normalization[153] tasks, since the adversarial penalty on the generator model is able to encourage better results than the generator alone.

There are also many limitations that still plague machine learning as a whole. First is the fact that deep learning requires incredibly large datasets to learn from[9]. Within the medical domain, medical images are expensive to create and difficult to obtain and use. This problem is compounded for supervised deep learning methods, which require not only the original data, but also the corresponding ground truths that they are attempting to predict[28]. In most applications, the creation of these ground truths falls upon researchers who must annotate, segment, label, and classify all the images in the large dataset. If these labeled datasets are not sufficiently large, the models will either fail to converge meaningfully or will learn to overfit the small dataset, making them unsuitable for application.

Further complicating this is the need for variation in the dataset[9]. Generalizability of models is important because many things can cause batch effects in the images, ranging from obvious things like the patient, operator, and laboratory to more trivial things like the weather and time of day[126]. Despite how important size and variation are to model training, there is no good method for estimating what is required ahead of time. Finally, even though models are trained using a provided ground truth, we must ask how we really define ground truth, whether the ground truths are adequate, and how much subjectivity we will accept in our definitions of ground truth. As previously discussed, even simple tasks like cell

counting produce various results between experts[111, 59], so when such manually labeled results are used as ground truth, the models will be subject to the data used for training. This same principle applies to other objectives such as feature extraction where there might not be easily identifiable/agreed upon truths that the model is trying to predict. Although deep learning models are powerful, so long as they are reliant on flawed human input for learning, they will be restricted in the scope of what they can achieve.

1.5 Dissertation contributions

In the following chapters, I will address several of the biomedical applications and limitations of the deep learning architectures described above. In chapter 2, I discuss the limitations of human annotators for segmenting cancer features from whole slide images for which there is no stain. Furthermore, within the chapter, I discuss the limitations of staining, thresholding, and normalization methods used for such tasks. I propose a UNet-based ensemble method with intermediate normalization steps called VIsual Semantic Tissue Analysis (VISTA)[126], which I show performs tissue segmentation in a fraction of the time and can guide researchers toward improved annotations.

The contents of chapter 2 are adapted from works listed below in chronological order:

- Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Joe W. Gray, John Muschler, Young Hwan Chang (2020, June 19). *Utilizing Deep Learning to Enhance and Accelerate Pancreatic Disease Quantification in Murine*

Cohorts [Conference poster]. Brenden-Colson Center, Portland, OR, United States.

- Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe W. Gray, John Muschler, and Young Hwan Chang. "Vista: Visual semantic tissue analysis for pancreatic disease quantification in murine cohorts". In: *Scientific Reports* 10.1 (2020). DOI: 10.1038/s41598-020-78061-3

- Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe W. Gray, John Muschler, Young Hwan Chang (2019). *VISTA: Visual semantic tissue analysis for pancreatic disease quantification in murine cohorts [Conference poster]*. Machine Learning for Health Workshop, Portland, OR, United States.

- Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe Gray, John Muschler, and Young Hwan Chang. "Abstract PO-014: VISTA: VIsual Semantic Tissue Analysis for pancreatic disease quantification in murine cohorts". In: *Cancer Research* 81.22 Supplement (2021), PO–014–PO–014. ISSN: 0008-5472. DOI: 10.1158/1538-7445.PANCA21-PO-014. eprint: https://cancerres.aacrjournals.org/content. URL: https://cancerres.aacrjournals.org/content/81/22_Supplement/PO-014

In chapter 3, I demonstrate the limitations of current VAE architectures for extracting features from single cell imaging data. I propose a novel architecture (the Multi-Encoder Variational AutoEncoder (ME-VAE)[124]) which attempts to overcome the noisy and biologically irrelevant transformational information present in

single cell images. I compare the tool to state-of-the-art methods and show that it improves downstream analysis via our ability to cluster cell types, extract novel features, and integrate with other modalities.

The contents of chapter 3 are adapted from works listed below in chronological order:

- Luke Ternes, Joe W. Gray, Laura Heiser, and Young Hwan Chang (2020, December 14). *Feature Controlled Variational Autoencoder for Single Cell Image Analysis [Conference poster]*. Learning Meaningful Representations of Life, Virtual. https://www.lmrl.org/posters2020
- Luke Ternes, Joe W. Gray, Laura Heiser, and Young Hwan Chang (2021, March). *Feature Controlled Variational Autoencoder for Single Cell Image Analysis [Conference presentation]*. CSBC / PS-ON Image Analysis Working Group, Virtual.
- Luke Ternes, Joe W. Gray, Laura Heiser, and Young Hwan Chang (2021, May 14). *ME-VAE:Multi-Encoder Variational AutoEncoder for Controlling Multiple Transformational Features in Single Cell Image Analysis [Conference poster]*. Human Tumor Atlas Network: Face2Face, Virtual.
- Luke Ternes, Joe W. Gray, Laura Heiser, and Young Hwan Chang (2021, July 25-30). *ME-VAE: Multi-Encoder Variational AutoEncoder for Controlling Multiple Transformational Features in Single Cell Image Analysis [Conference poster]*. International Society for Computational Biology, Virtual. https://www.youtube.com/watch?v=fGgVYV0nBoA

- Luke Ternes, Mark Dana, Marilyne Labrie, Gordon Mills, Joe W. Gray, Laura Heiser, and Young Hwan Chang (2020, Nov 23-24). *ME-VAE: Multi-Encoder Variational AutoEncoder for Controlling Multiple Transformational Features in Single Cell Image Analysis [Conference poster]*. MLCB: Machine Learning in Computational Biology, Virtual.
- Luke Ternes, Mark Dane, Sean Gross, Marilyne Labrie, Gordon Mills, Joe Gray, Laura Heiser, and Young Hwan Chang. "ME-vae: Multi-encoder variational AutoEncoder for controlling multiple transformational features in single cell image analysis". In: (2021). DOI: 10.1101/2021.04.22.441005
- Luke Ternes, Mark Dana, Marilyne Labrie, Gordon Mills, Joe W. Gray, Laura Heiser, and Young Hwan Chang (2022, Jan 3-7). *Extracting more biologically relevant features from multiplexed imaging with a Multi-Encoder Variational AutoEncoder [Conference presentation]*. Pacific Symposium on Biocomputing, Waimea, HI, United States.

In chapter 4, I work to mitigate the burden on the multiplex imaging pipeline through three tasks: stain prediction in 3-dimensional tissue volumes, representative region-of-interest identification via optimization, and quantitative multiplex image panel reduction. I propose the use of a previously established stain prediction algorithm (SHIFT)[133] for the prediction and propagation of multiplex staining throughout a 3-dimensional tissue volume. I demonstrate the use of a GAN-based XAE architecture and convex optimization function for discovering representative regions-of-interest (ROIs) which can characterize whole slide images without the need to stain or analyze the whole section. Finally, I evaluate

several methods for selecting an ideal/reduced panel for multiplex imaging that maximizes the amount of information retained while eliminating stains whose information can be captured from markers elsewhere within the panel.

The contents of chapter 4 are adapted from works listed below in chronological order:

- Luke Ternes, Erik Burlingame, Jia-Ren Lin, Yu-An Chen, Eun Na Kim, Joe W. Gray, Sandro Santagata, Peter Sorger, and Young Hwan Chang (2021, Nov 18-19). *3D reconstruction of whole-slide multiplex tissue imaging and optimized ROI selection*. (paper in preparation).
- Luke Ternes, Erik Burlingame, Jia-Ren Lin, Yu-An Chen, Eun Na Kim, Joe W. Gray, Sandro Santagata, Peter Sorger, and Young Hwan Chang (2021, Nov 18-19). *3D reconstruction of whole-slide multiplex tissue imaging and optimized ROI selection with deep learning [Conference poster]*. Human Tumor Atlas Network Face2Face, Virtual.

3D multiplexed tissue imaging reconstruction and optimized region-of-interest selection through deep learning model of channels embedding.

1.6 Other contributions

Other contributions, manuscripts, and publications completed during my doctoral studies have been omitted to maintain a clear focus in this dissertation. These works are listed below in chronological order:

- Luke Ternes, Caitlin Mills, Kartik Subramanian, Yunguan Wang, Clarence Yapp, Sean Gross, LINCS MCF10A Consortium, Joe W. Gray, Peter Sorger, Laura Heiser, and Young Hwan Chang, (2019, Oct 2-4). *The Temporal Dynamics of Ligand Treated MCF10A Cells Using Cyclic Immunofluorescent Imaging Data [Conference poster]*. Allen Institute BioImage Informatics Symposium, Seattle, WA, United States.

- Luke Ternes, (2020, Jan 9-10). *Recursive Segmentation Refinement Without Manual Annotations [Conference presentation]*. CSBC / PS-ON Image Analysis Working Group, Seattle, WA, United States.

- Luke Ternes, Guillaume Thibault, Joe W. Gray, Young Hwan Chang, (2020, Apr 3-7). *Iterative deep learning based segmentation on cyclic immunofluorescence imaging by using recursive refinement [Conference presentation]*. IEEE International Symposium on Biomedical Imaging, Iowa City, IA, United States.

- Juan Carlos Vizcarra, Erik A. Burlingame, Clemens B. Hug, Yury Goltsev, Brian S. White, Darren R. Tyson, and Artem Sokolov. "A community-based approach to image analysis of cells, tissues and tumors". In: *Computerized Medical Imaging and Graphics* 95 (2021), p. 102013. DOI: 10.1016/j.compmedimag.2021.102013

- Orit Rozenblatt-Rosen *et al.* "The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution". In: *Cell* 181.2 (2020), pp. 236–249. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2020.03.053. URL: https://www.sciencedirect.com/science/article/pii/S0092867420303469

Chapter 2

Reproducible segmentation of nuanced cancer features without annotator bias

A frog in a well cannot discuss the ocean,
because he is limited by the size of his well.

Readings from Chuang Tzu

2.1 Abstract

Mechanistic disease progression studies using animal models require objective and quantifiable assessment of tissue pathology. Currently quantification relies heavily on staining methods which can be expensive, labor/time-intensive, inconsistent across laboratories and batch, and produce uneven staining that is prone to misinterpretation and investigator bias. I developed an automated segmentation

tool (VISTA) utilizing deep learning for rapid and objective quantification of histologic features at the pixel-level, relying solely on hematoxylin and eosin stained pancreatic tissue sections. The tool segments normal acinar structures, the ductal phenotype of acinar-to-ductal metaplasia (ADM), and dysplasia with Dice coefficients of 0.79, 0.70, and 0.79, respectively. To deal with inaccurate pixelwise manual annotations, prediction accuracy was also evaluated against biological truth using immunostaining mean structural similarity indexes (SSIM) of 0.925 and 0.920 for amylase and pan-keratin respectively. Our tool's disease area quantifications were correlated to the quantifications of immunostaining markers (DAPI, amylase, and pan-keratin; Spearman correlation score= 0.86, 0.97, and 0.92) in unseen dataset (n=25). Moreover, our tool distinguishes ADM from dysplasia, which are not reliably distinguished with immunostaining, and demonstrates generalizability across murine cohorts with pancreatic disease. I quantified the changes in histologic feature abundance for murine cohorts with oncogenic Kras-driven disease, and the predictions fit biological expectations, showing stromal expansion, a reduction of normal acinar tissue, and an increase in both ADM and dysplasia as disease progresses. Our tool promises to accelerate and improve the quantification of pancreatic disease in animal studies and become a unifying quantification tool across laboratories.

2.2 Introduction

Advances in deep learning technologies are creating opportunities for the rapid and objective assessment of both normal tissue and pathologic processes in biologic specimens. Computer-aided interrogation of medical imaging is being applied to accelerate and improve diagnosis in human patients[17, 74, 20, 110]. Similarly, deep learning technologies can greatly improve analyses in animal disease models which require the measurement of disease progression in large numbers of tissue samples resulting either from pharmacological or genetic manipulations. The extensive and growing use of murine models in disease studies creates a significant need for tissue assessment methods that are rapid, objective and quantifiable in order to permit statistically validated disease measurements among animal cohorts, free of technical variability and investigator bias.

The challenge of objective quantification of tissue changes among animal cohorts is significant. Evaluation of tissue by either histochemical stains or antigen-specific immunohistochemistry offers distinct and sometimes overlapping information, but both have limitations. Hematoxylin and eosin (H&E) staining is a rapid, reliable and inexpensive method; however, lack of molecular specificity and requirement for manual segmentation have, thus far, limited its use for extraction of quantifiable data. Consequently, disease assessments by H&E staining are typically qualitative and vulnerable to inter-observer variation and bias[31, 149, 27]. Immunohistochemical stains offer a degree of specificity, but immunostaining can be labor- and time-intensive, expensive and results are often challenging to objectively quantify over broad tissue regions. In addition, tissue features of

interest are not always cleanly distinguishable by immunostaining markers, and so tissue assessments can be limited by reliance on the molecular specificity of antibodies.

Here I develop and validate deep learning approaches that enable the rapid, reliable, and automated quantification of disease progression over large tissue areas, solely based on H&E staining, using murine models of pancreatic cancer progression and pancreatitis. Murine models of pancreatic cancer were chosen as they have proven useful for mechanistic investigations of pancreatic cancer progression, modeling well the human disease both genetically and phenotypically, particularly during the evolution of pre-cancerous lesions[45, 46]. The murine models have produced an explosion of studies including pre-clinical drug tests and evaluation of additional genetic perturbations that expose tumor-suppressing and tumor-promoting disease modifiers[148, 132, 141].

The early stages of pancreatic cancer evolution are well described in the mouse models[45, 46]. The normal pancreas consists predominantly of acinar and ductal epithelial cells forming the exocrine compartment, along with islet cells of the endocrine compartment, vasculature and the varied fibroblasts of the stromal compartment. The earliest stages of oncogene-induced pre-cancer evolution are marked by an expansion of ductal cells or by the conversion of the acinar cells to a ductal phenotype in an adaptive process known as acinar-to ductal metaplasia (ADM)[118]. ADM is also characteristic of acute and chronic pancreatitis, inflammatory conditions that can predispose to cancer[118]. The next stage in cancer evolution is the development of low-grade dysplasia, also referred to as pancreatic intraepithelial neoplasias (PanINs 1 and 2). Low-grade dysplasia is a

pre-invasive neoplasia that can evolve to high-grade dysplasia (PanIN 3) and then progress to invasive pancreatic ductal adenocarcinoma (PDAC)[97]. Both ADM and dysplasia are accompanied by a prominent stromal reaction and immune cell infiltrate[118]. The stages of ADM and dysplasia evolution are believed to encompass a long phase of pre-cancer evolution that is a valuable window for early intervention[97].

Within this work, I describe the model training workflow and application of deep learning on H&E stained samples of murine pre-cancerous lesions, segmenting the normal acini, the ductal phenotype of ADM, and dysplasia. With the rapid growth of computer vision, more specifically deep learning, novel image analysis architectures have been developed for accessing image information that is not readily observed through traditional methods. Several research groups have worked towards inter-modality image translation and have developed tools that attempt to convert medical images such as H&E stained tissue and brightfield microscopy to more detailed ones such as fluorescent immunostains[16, 23, 91, 61]. The target of such models has been the direct translation of stain intensities for the purpose of constructing entirely new images. Our developed tool seeks to go further, predicting binarized masks of positive staining area and augment immunostaining by segmenting key histologic features that current stains cannot reliably differentiate.

Results presented here demonstrate a well validated segmentation tool that can automatically, rapidly, and objectively quantify pancreatic tissue and disease progression in mice, relying solely on easily replicated and low-cost H&E staining of whole pancreas tissue sections, free of experimental variability and investigator

bias. Our work provides a tool that is immediately applicable to the improvement and acceleration of pancreatic disease studies in animal cohorts, and provides workflows for similar tool development in other disease models. Moreover, the ease of use and availability allows for this tool to be a common thread for comparing different studies performed throughout the world.

2.3 Results

In order to predict the histologic feature distributions and immunofluorescent stain positivity in murine pancreatic pre-cancerous tissues, several UNet models[102] were trained using intensity normalized H&E image tiles paired with annotated ground truth tiles (Figure 2.1). All pancreas tissue sections in training, validation, and testing sets were stained with H&E (Table 2.1). First testing was conducted by evaluating spatial overlap of predictions and expert annotations for normal acinar, ADM, and dysplasia. A second test was performed by correlating predictions to binarized immunofluorescence staining (IF): amylase (AMY), labeling normal acini, pan-keratin (panK), labeling primarily the oncogenic Kras-transformed epithelial population, and DAPI, labeling all nuclei. A third test was performed qualitatively analyzing predictions in pancreatitis and normal samples, and comparing to biological expectations.

To ensure that UNet[102] model predictions were able to generalize well and overcome staining differences within and between tissue sections, Dice Coefficient were calculated comparing model predictions to expert annotations made in Cytomine[82] after training with several different normalization techniques

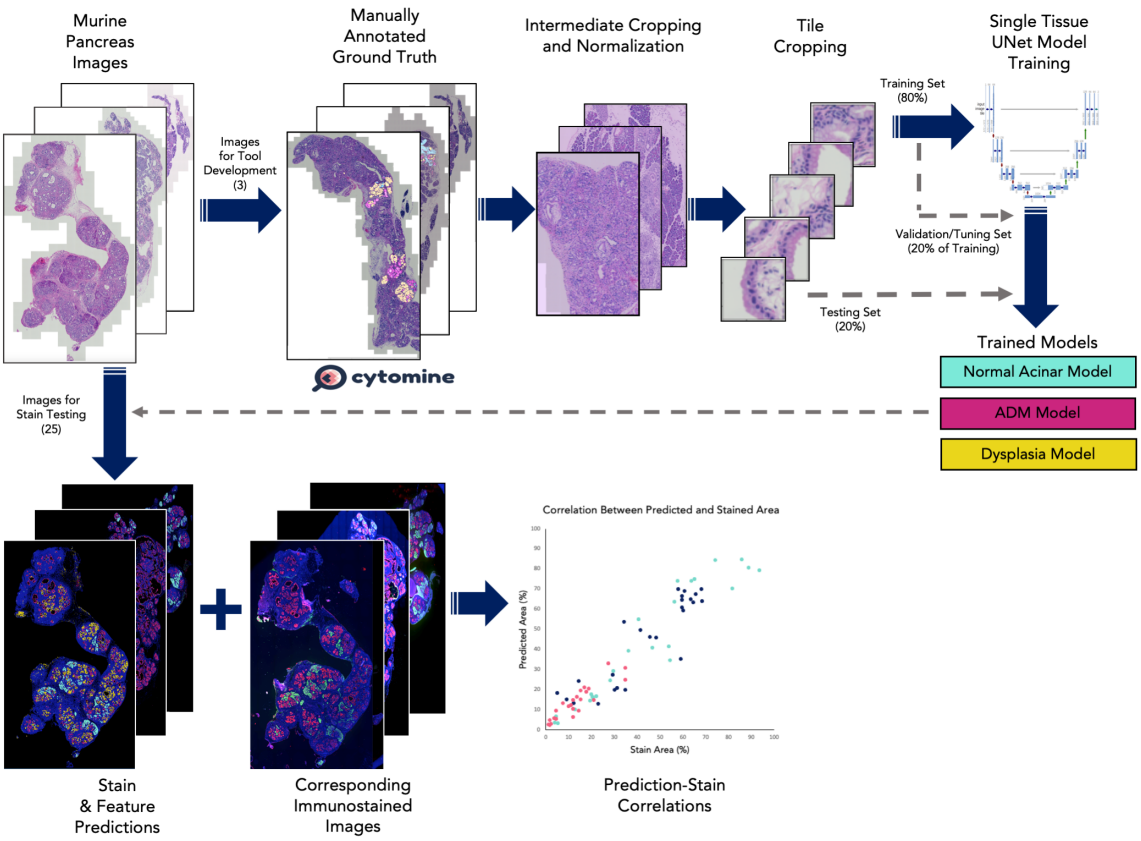


FIGURE 2.1: A subset of murine pancreas H&E images were annotated by three experts in Cytomine[82]. The images and their annotations were cropped and normalized at intermediate intervals, and these intermediate crops were then tiled into images that can be fed into a UNet architecture[102]. 80% of tiles were used for training and validation, and 20% of tiles were used for testing. A model was trained for each histologic feature label. The best models were chosen and used to predict stain and feature distributions on unseen H&E images. These predictions were then correlated with the stained image counterparts to determine model accuracy.

(Reinhard[98], Vahadane[134], and Macenko[76]). As observed in Table 2.2, the models implementing Reinhard normalization achieved better scores on average, relative to Vahadane and Macenko. Furthermore, the models achieved the best scores when the normalization process was applied on intermediate crops rather than across the whole image. This is because staining can be uneven within a single section, and normalizing crops helps to overcome these differences in

	Sample Size	H&E	IF	Annotations	Used For
KC (2 months)	12	x	x		- IF Correlation -Area evaluations in pre-cancer histopathology
KC (5 months)	16	x	x (n=13)	x (n=3)	- Training and validation - Dice evaluation (H&E-based prediction vs Annotations) - IF Spearman correlations - Area evaluations in pre-cancer histopathology - SSIM evaluation (H&E-based prediction vs IF stain) - Synthetic stain generalizability
Induced Pancreatitis	6	x			- New tissue generalizability
Normal Tissue	3	x			- New tissue generalizability

TABLE 2.1: Datasets Used

intensity; whereas, normalizing across a whole section only helps overcome differences between images.

The models were trained using 80% of the training dataset, and 20% of that training dataset was held out for cross-validation to evaluate and tune the models' performance with unbiased data. The unseen labeled test data, comprising 20% of the annotated dataset was used for final evaluation of the models. The best models yielded Dice Coefficients of $\sim$0.79, $\sim$0.70, and $\sim$0.79 on the hold-out set for normal acinar tissue, ADM, and dysplastic features, respectively (Table 2.2). The segmentations match the expert annotations with a high degree of qualitative accuracy (Figure 2.2a). The reason that the models' Dice scores are lower than expected from successful models is because the models actually refined approximations in the experts' annotations leading to discrepancies between prediction and annotation (Figure 2.2b). Due to the limitations of the annotation method used, entire lesions (including empty lumina, mixed morphologies (Figure 2.3e), and additional

negative space) were labeled as one type of tissue (i.e., ADM or dysplasia). The models, however, accurately differentiate between the tissue types within a lesion and avoid labeling lumina. Despite the predicted histology labels being biologically correct, they differ than the experts' manual annotations, resulting in a negative impact on the measured Dice Coefficients. This means that the performance of the models is actually greater than reported by Dice due to inconsistent biological error in the annotations which were used for model evaluation.

Normalization Method	Metric	Normal Acinar	ADM	Dysplasia
Reinhard Normalization of Intermediate Crops	Dice	**0.78691**	**0.70239**	**0.79403**
	BCE	0.16131	0.17112	0.22374
Reinhard Normalization[98]	Dice	0.71750	0.60303	0.76210
	BCE	0.20561	0.16635	0.21966
Vahadane Normalization[134]	Dice	0.69311	0.58241	0.73684
	BCE	0.20753	0.18726	0.24471
Macenko Normalization[76]	Dice	0.70686	0.56660	0.77210
	BCE	0.21784	0.18370	0.19711

TABLE 2.2: Evaluation of Model Performances

To test the accuracy of the trained models further, a comparison was made between quantified model predictions and a second unseen test set of immunostained images that have been binarized. Quantification of the tissue area occupied by normal acinar cell and transformed pancreatic epithelial cells was achieved by immunostaining for amylase and panK, respectively, with DAPI staining of nuclei used to detect all cellular regions. The comparable calculation was then made using tool predictions on adjacent H&E stained tissue sections. For the tool prediction, ADM and dysplasia predictions were grouped into the panK stain because panK immunostaining does not distinguish ADM and neoplastic tissues. Because stain area is specific and more biologically targeted than the rough annotations that incorporate empty lumens and mislabeled features, the models' immunostain

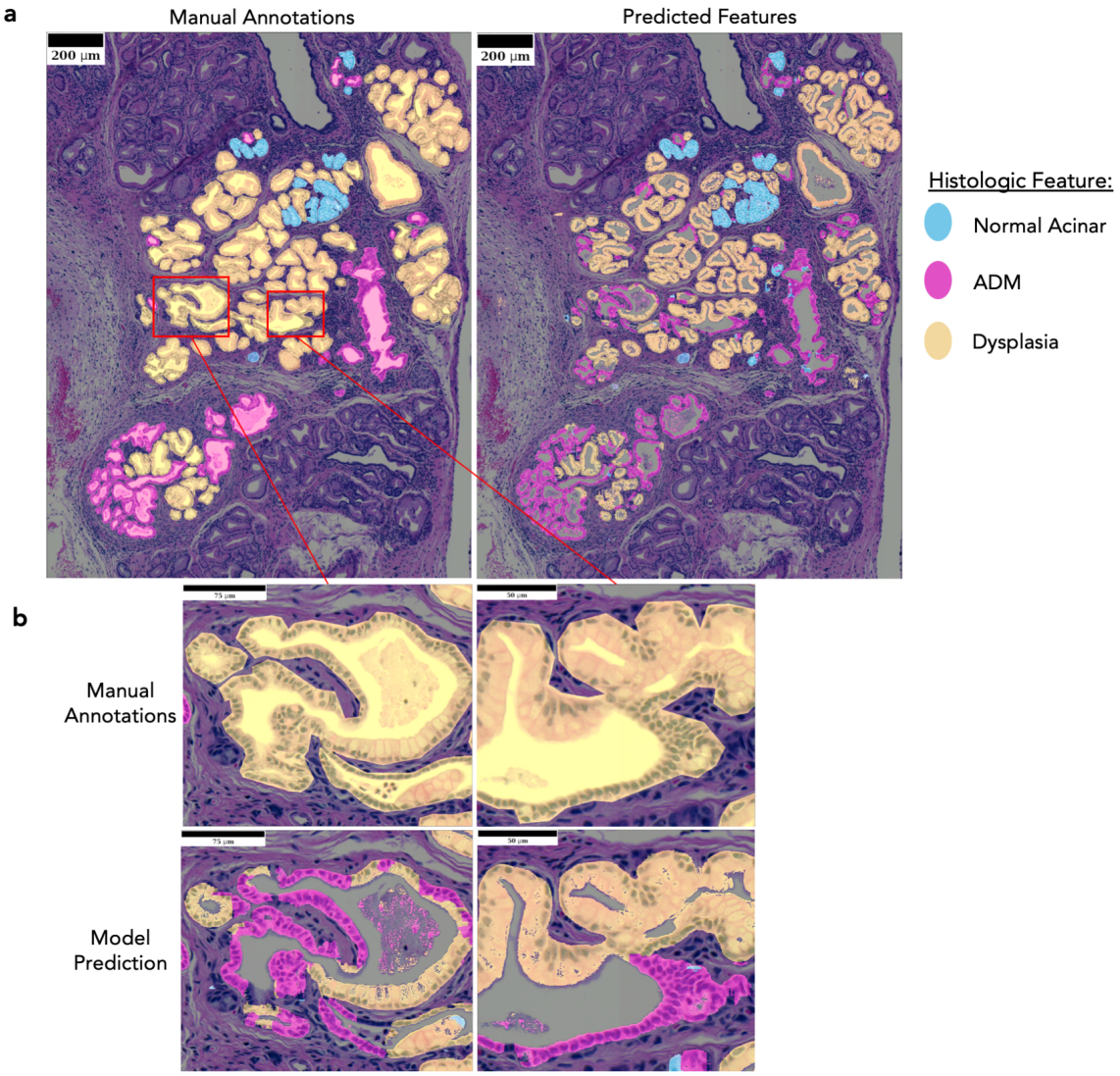


FIGURE 2.2: a) Model Predictions closely align with the manually annotated ground truth regions that was used for training. b) Close inspection of the ducts shows consistent discrepancies regarding the lumen and split histologic features within single ducts. Manual annotations were made by circling whole ducts, but the models' predictions are actually more reflective of biology, wherein, stain does not mark for the lumen. The Predictions can also distinguish histologic features differences that the manual annotations combined.

Spearman correlation scores are much more reflective of their overall accuracy and

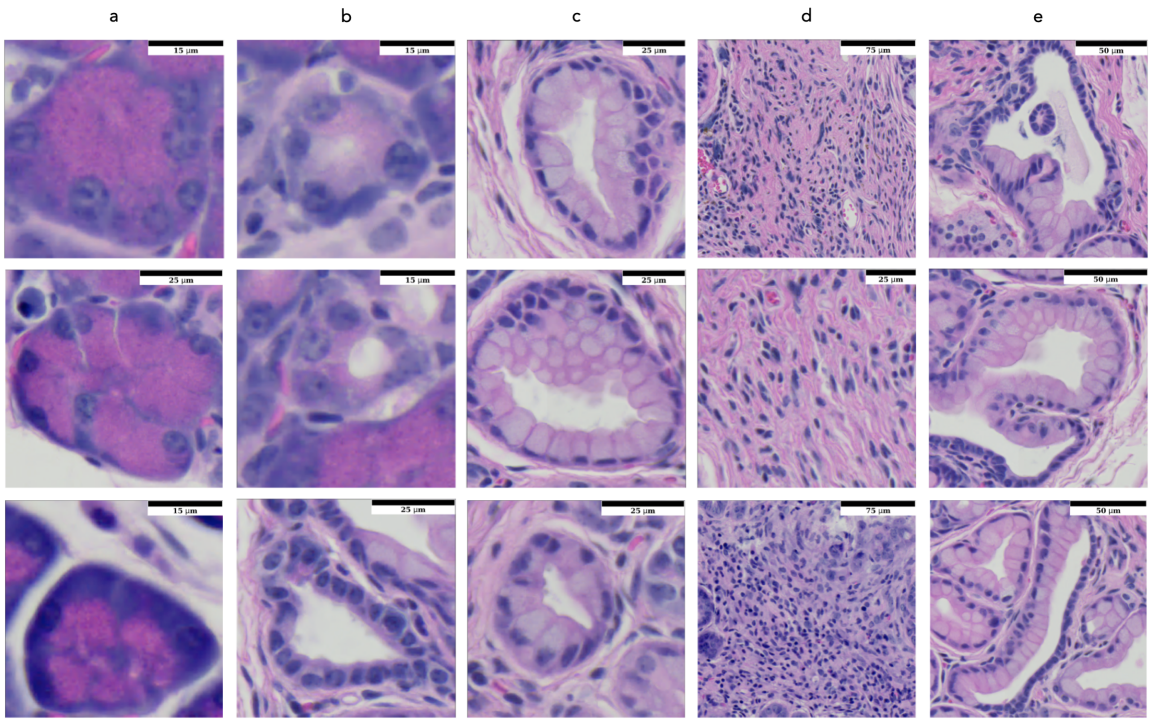


FIGURE 2.3: The most prominent histologic features of the pre-cancerous pancreas are the normal acinar epithelial cells (a), the ductal structures resulting primarily from ADM (b), dysplasias (c), and the inflammatory and ECM-rich stroma (d). The normal acini are marked by a thick and darkly stained cytoplasm. ADM is distinguished by a diminished stain, reduced cytoplasm and frequent appearance of a ductal lumen. Low-grade dysplasias are distinguished primarily by an enlargement of the cytoplasm that is lightly stained and correlates with enhanced intracellular mucin production. Dysplasia can exhibit a hybrid appearance of flattened duct-like cells and thickened mucin-rich neoplasia (e). The reactive stroma is marked by dense ECM, spindle like fibroblasts, and inflammatory cells.

sensitivity. When the prediction masks are compared qualitatively and quantitatively to the stained images, the models are able to predict the spatial localization of the immunostaining (Figure 2.4a and Figure 2.5). Prediction accuracy was evaluated against biological truth using immunostaining and structural similarity (SSIM) (Figure 2.5), in addition to the area correlations (Figure 2.4). SSIM was chosen as our metric to evaluate against because it would be more robust than Dice against differences between serial sections. Note that H&E and IF stained samples

were acquired from adjacent serial sections.

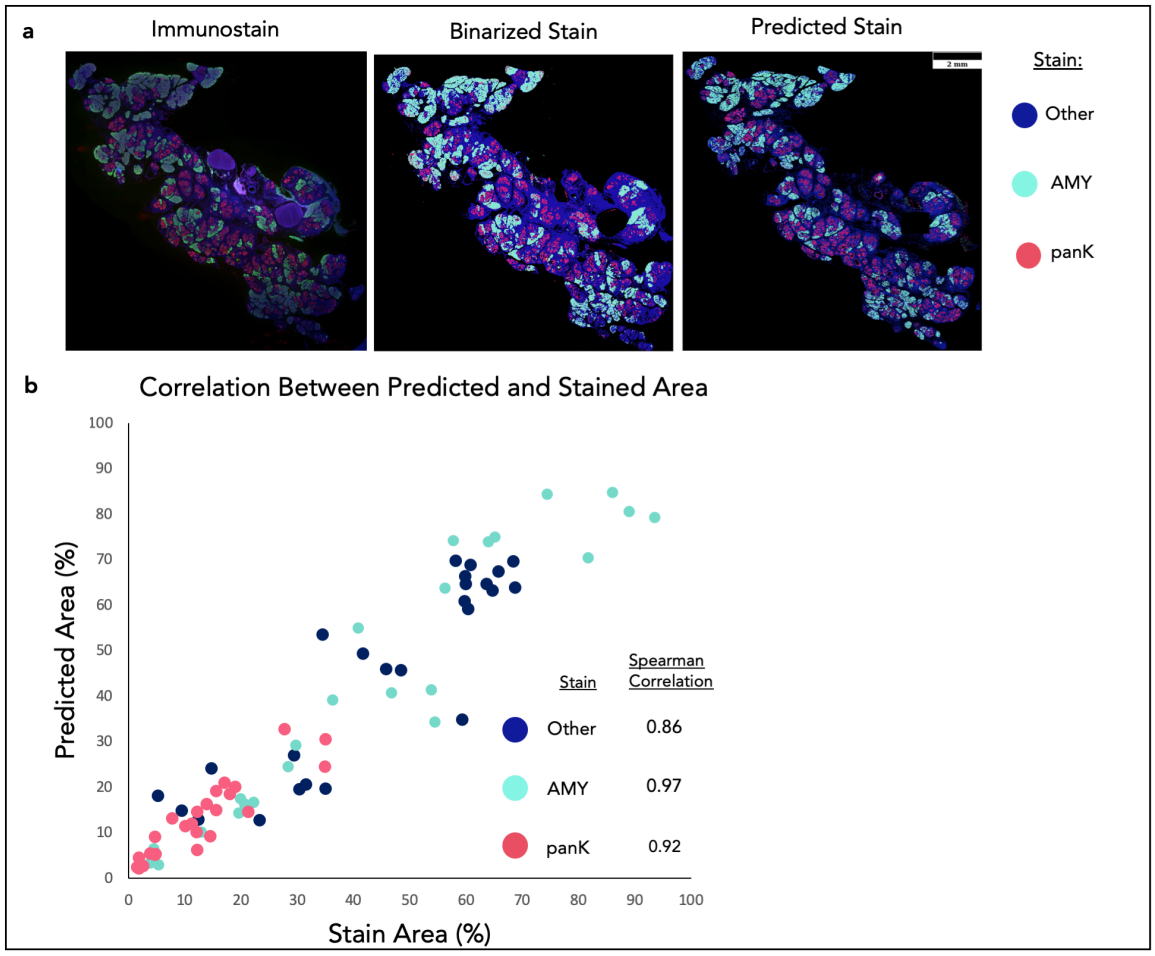


FIGURE 2.4: a) Stain masks and predicted segmentation masks are qualitatively highly similar. Differences can be seen in the high-level architecture of the tissues, which is indicative of the fact that the predictions were made from serial sections to the stains. There are also dim regions of the stained image that are lost from the global thresholding technique. These regions are successfully captured by the models. "Other" stain is the DAPI stain minus regions overlapping with AMY and panK. b) Correlations were made by comparing the percent of area coverage for each stain mask. The high Spearman correlations illustrate the models' ability to replicate straining using only H&E images. These regions are successfully captured by the models. "Other" stain is the DAPI stain minus regions overlapping with AMY and panK.

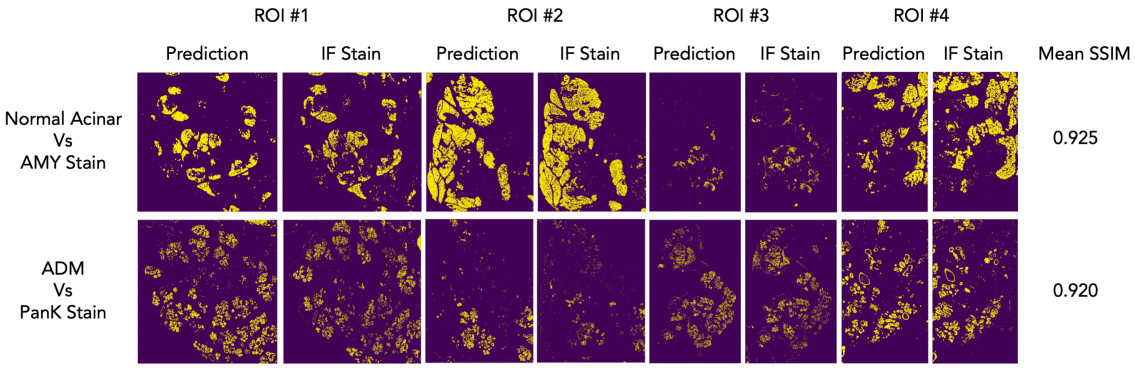


FIGURE 2.5: Binarized stain was obtained using thresholding and compared to the corresponding histological predictions for 4 ROIs using SSIM. Small Gaussian blurs were used to account for the fact that samples were taken in serial sections. The high SSIM score support the accuracy of the tool at predicting histological information.

There are minor differences between the immunostained and the predicted segmentations, which reflects slight tissue variations between the adjacent, but separate, sections used for H&E and IF staining. Quantitatively, three models also have high Spearman correlations (Figure 2.4b) with the immunostained sections despite these sections (n= 25) being unseen during training, with Spearman correlation values of 0.97, 0.92, and 0.86 for AMY, panK, and DAPI stained other tissue, respectively. These correlations are very strong, despite the assumption that the serial section have true correlation values of close to 1[21]. The good qualitative spatial localization and strong correlations validate that the models have been successfully trained and are capable of replicating known biological data.

Not only can these models replicate immunostaining data, they can extract more information than can be gained via immunostaining. In the second unseen testing dataset consisting of 25 IF/H&E image pairs, the panK immunostain labels both metaplasia and dysplasia, restricting the disease features that can be segmented.

The model predictions, however, can distinguish these features (Figure 2.6a). This allows for deeper and more nuanced quantification of disease progression than can be achieved by immunostaining alone. Across a whole section of unseen test tissue, it can be observed that each predicted feature corresponds with the correct morphology.

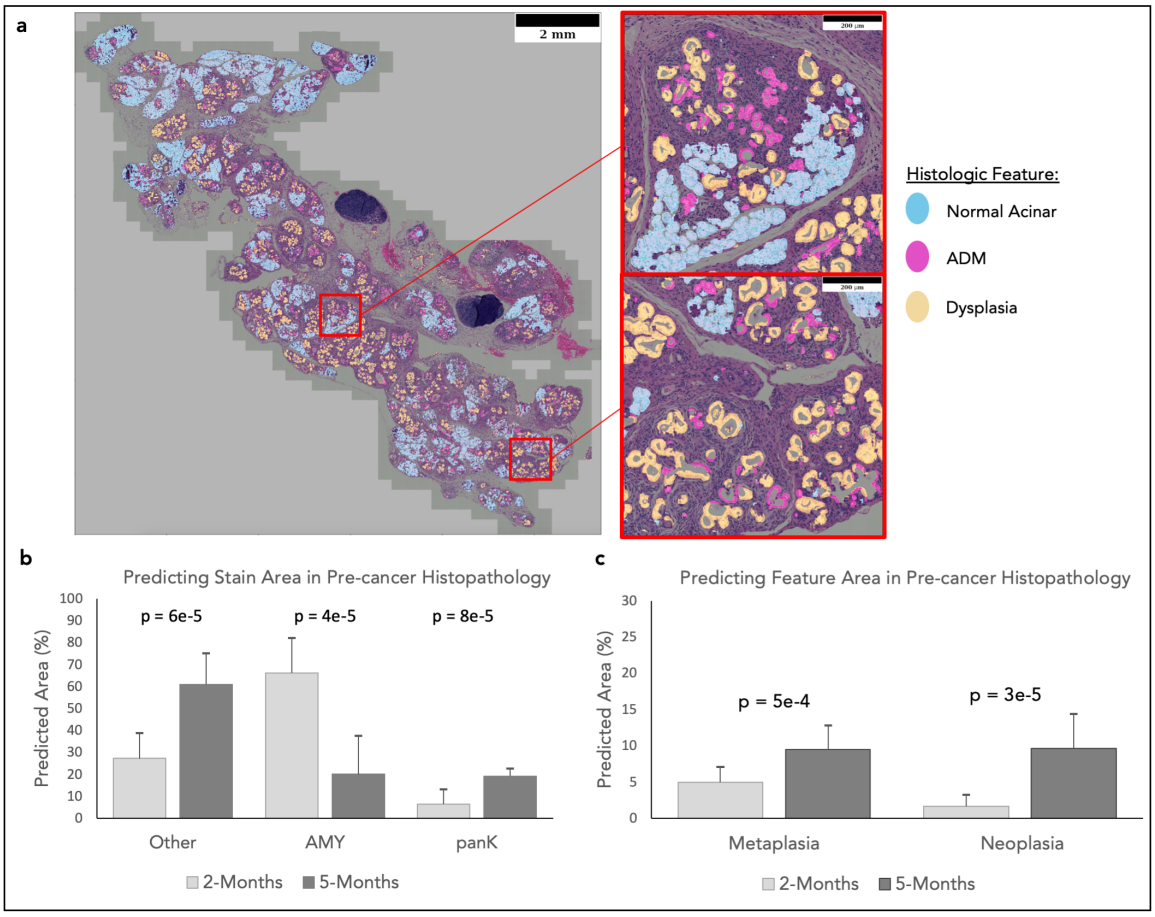


FIGURE 2.6: a) In test images the predicted histologic features visually align with what is expected from the H&E images. This shows the models' utility in discerning novel information regarding ductal features that cannot be detected via staining.
The models were used to predict the changes stain distributions b) and cancer histologic features c) in murine models with induced cancer. Predictions show significant changes in all stains and features between time points, and quantifies specific features that were not discernable in immunostaining alone. Mann-Whitney U test was used to test for statistical analyses.

Because this process of prediction is deterministic, it is also a faster and less biased than manually annotating histologic features, and less expensive and less error-prone than immunostaining (Figure 2.7). Standard binarization of whole slide IF stains often leaves dimmer regions of the tissue with inaccurate predictions of stain positivity. This process of setting a threshold for stain binarization is also a subjective process that will have different results depending on the expert looking at the image, and performing regional thresholding throughout the image demands more time and introduces more thresholds that can be biased by the evaluator. By comparison, the trained models are deterministic and are able to overcome staining differences in a consistent manner. Furthermore, the process of staining an IF section takes two days following standard protocol, with additional time spent image processing and binarizing the image afterwards. By comparison, the deep learning models take less than an hour depending on section size and graphics process unit performance. Human annotation of the data is even slower, taking days to weeks for a single section and can have high variability between annotators. In addition, it can be difficult to get access to an expert with pathology certification necessary for differentiating the morphologies.

Using the tissue sections from the second unseen testing dataset isolated from $P48^{+/Cre}; LSL-KRAS^{G12D}$ (KC) mice at 2 and 5 months of age (n=12, n=13), the model was able to quantify tissue changes reflecting disease progression by predicting immunostain from H&E stain images (Figure 2.6b/c). The observed age-dependent transitions from normal acinar to ADM and dysplasia, and the increase in other tissue area (DAPI stained), is consistent with biological expectations, illustrating the practical, objective use of this tool to quantitatively assess pre-cancerous

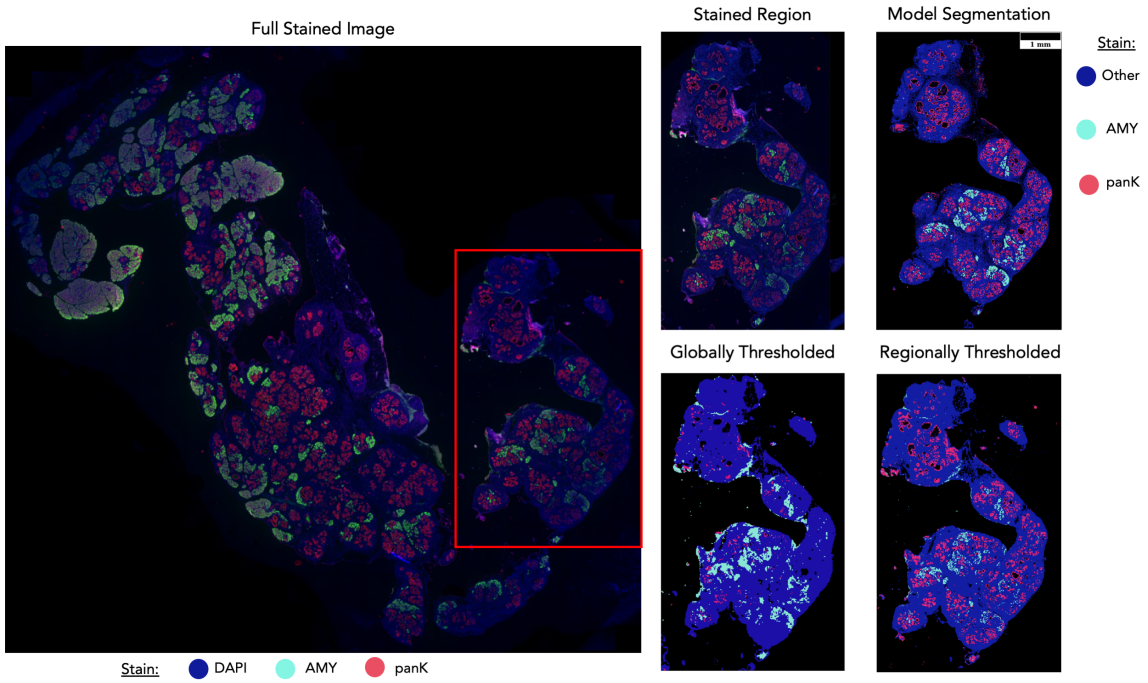


FIGURE 2.7: The quality of the full stained image varies region to region, as some regions have dimmer staining than others. Because of this uneven staining quality, a single global threshold will not accurately represent true positives and negatives because dimmer regions will be neglected. When regions are thresholded independently, the quality of the segmentation masks improves; however, even regional dim spots are still dropped from the segmentations. The developed models, however, are able to overcome this limitation because it utilizes H&E images and is able to analyze the histologic features beyond just the intensity of the stain. "Other" stain is the DAPI stain minus regions overlapping with AMY and panK.

disease development.

To test the models' robustness and generalizability, I evaluated images from pancreata exhibiting histopathology associated with acute pancreatitis instead of histopathology induced uniquely by oncogenesis. Acute pancreatitis is characterized by prominent ADM and an inflammatory stromal response, but does not promote neoplastic lesions[118]. Acute pancreatitis was induced in mice by injection of the pro-inflammatory agent caerulein[118], then tissue sections

exhibiting acute pancreatitis or normal pancreas (n=6, n=3 respectively) were analyzed by the model (Figure 2.8). This was performed on a third test dataset not seen by the models during training. Because neither annotations nor stains exist for this third dataset, model prediction localizations were evaluated qualitatively. Despite not being trained to analyze the particular disease states of pancreatitis, the models were able to accurately label pancreatitis features (i.e. ADM) with minimal error, regardless of whether the ADM was sporadic or clustered within the tissue (Figure 2.8a). The model's quantified tissue assessments show the significant presence of ADM by pixel area in the pancreatitis samples compared to normal tissues, which matches biological expectations. The near-absence of significant ADM and dysplasia in normal pancreas samples is also consistent with expectations, as is the near-absence of dysplasia in the pancreatitis samples (Figure 2.8b). The small quantities of ADM and dysplasia predictions in the normal tissues are errors introduced primarily by pixel level noise and are insignificant compared to the size of the samples. Within this dataset I do not see large heterogeneity in the histologic features across disease states, and as a result the model performs consistently across all disease states shown.

2.4 Discussion

The computational tool developed here is intended to augment and accelerate disease research performed in animal models by allowing for simple stain prediction and histologic feature labeling from H&E images without the need for expensive and time-consuming immunostaining and biased image interpretation. It can be

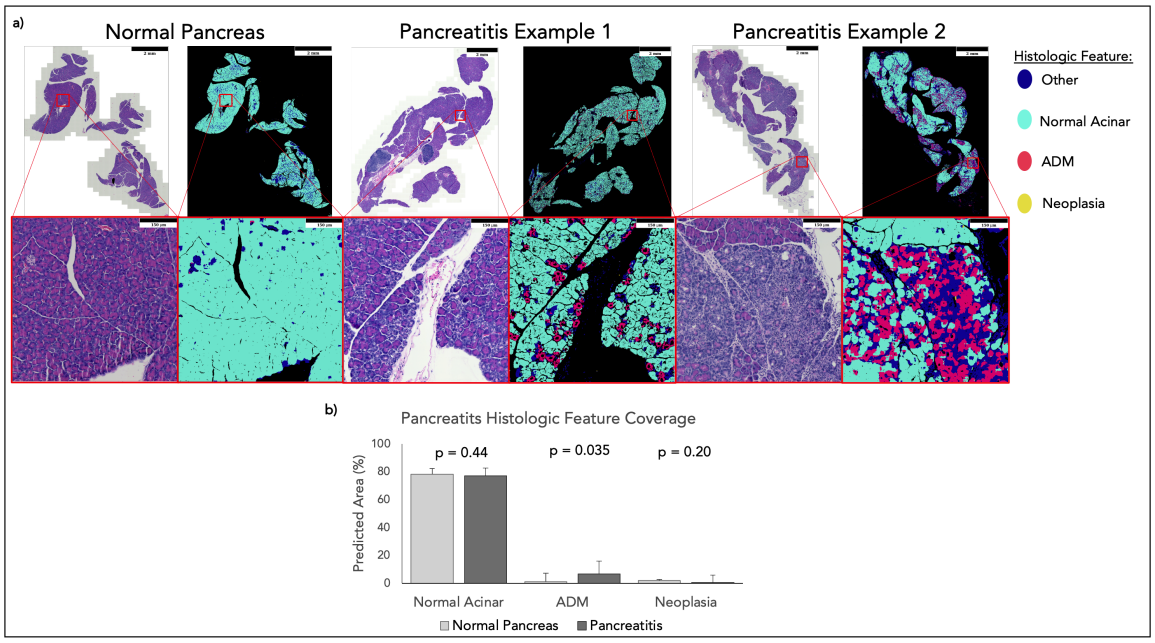


FIGURE 2.8: The model predicted histologic features match what in expected in both normal and pancreatitis samples. a) Predicted images show that tissue is dominated by normal acinar with pockets of clear ADM localization. In normal tissue ADM and dysplasia are sparse predictions comprised primarily of arbitrary single pixels, and in pancreatitis this is true for just dysplasia.
b) In normal tissues, ADM and dysplasia predictions are negligible, and in pancreatitis there is a significant spike in ADM coverage with negligible dysplasia. Mann-Whitney U test was used to test for statistical analyses. Erroneous predictions of ADM and dysplasia in these samples are primarily driven by noise.

used to both mark the localization of tissue features and quantitatively to measure the extent of disease based on multiple histologic features (Figure 2.3). Such rapid and unbiased quantification of disease states in animal models is critical to enabling efficient and accurate disease assessments among large study cohorts, as well as provide a common method to compare finding across different studies. The ability of this tool to accurately predict histologic features among 25 unseen pancreatic pre-cancer samples from multiple time points and 9 unseen samples comprising other disease states demonstrates the robustness of the models when

analyzing new datasets. The fact that the models generalize well, despite being trained with a relatively small dataset (Figure 2.9 and Table 2.3), illustrates the effectiveness of this workflow for tool development. Using this workflow (Figure 2.1) makes niche tool development plausible for small working groups that might have less access to the resources needed to produce large batches of annotated data. This pipeline is also faster, cheaper, and more generalizable than immunostaining, which can take days and be prone to investigator bias. This will allow working groups to digitally process many samples within hours instead of spending days immunostaining individual samples.

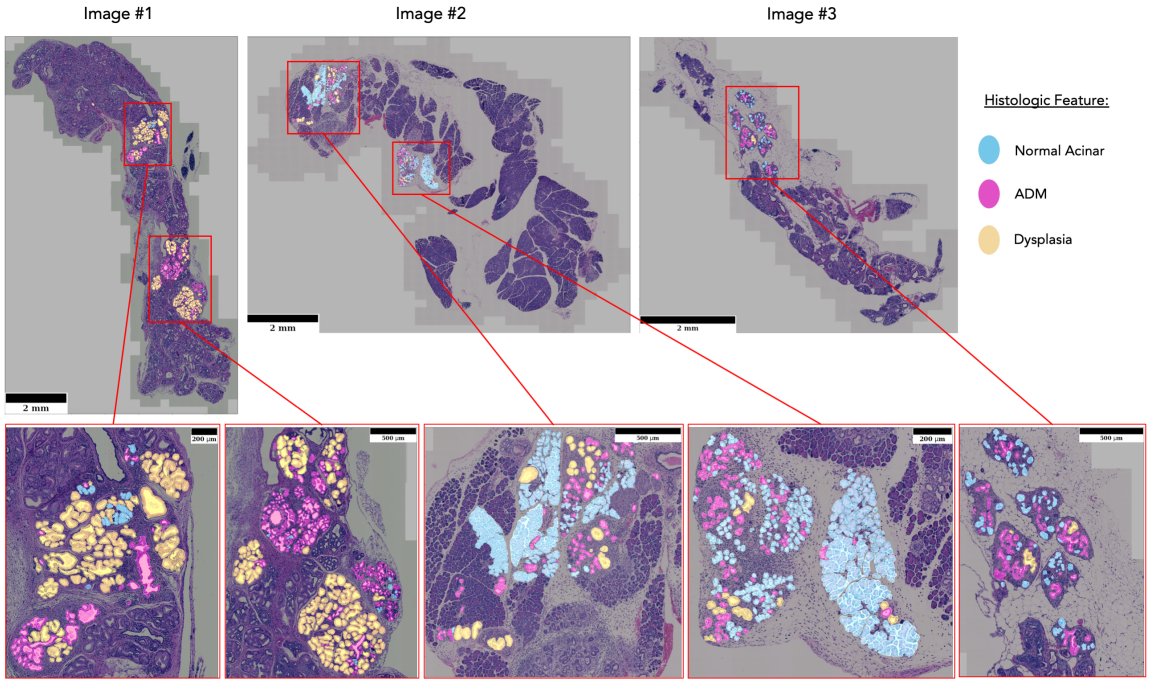


FIGURE 2.9: The annotated training, validation, and testing dataset was comprised of labels from 5 regions across three images. In the context of deep learning, where some models are trained on thousands of whole tissue sections, this is considered a small dataset. Despite only being trained on annotations from 3 images, however, the models are successfully generalizable to many more images with different staining qualities and levels of cancer development.

There have been many efforts to recreate advanced staining images using more common input modalities[16, 23, 91, 61], and although they are useful for visualizing potential stain and intensity distributions, the algorithms are limited to predicting staining patterns of existing markers. If the user wants to analyze specific biological features for which there is no specific stain; however, simple stain translation will not suffice. The tool created here, however, can create objective binary interpretations of H&E images that segment histologic features of developing pancreatic cancer for which there is no reliable conventional immunostain. Although this methodology uses a conventional UNet architecture[102], which is a standard method used for segmentation in deep learning, the work presented here is a novel and useful application of this technology for studies into the development, progression, and mechanisms of pancreatic pre-cancers. The tools allows researcher to distinguish the developmental histological cancer features of ADM and PanIN lesions, which have seen few applications before and are important for quantifying disease progression. These features, as previously described, are not easily distinguished by any other methods besides manual annotation. Previous studies have attempted to use computer-aided analyses for duct detection in pancreatic cancer[62], and although the results are good, they are limited in their scope and do not cover a range of subtly different features or early disease hallmarks such as ADM and dysplasia. This illustrates the capacity for modern deep learning methods to provide a broader range of information and perform more complex tasks with comparable accuracy.

Although this tool enables easy, rapid, and accurate binary stain prediction and feature labeling in the early stage disease models employed here, there are several

limitations to its predictive capacity. The most prominent source of error for the tool currently is the way it handles unlearned tissue types, such as lymph nodes, pancreatic islets, the desmoplastic stroma, and the occasional presence of neighboring gastrointestinal tissue. Lymph nodes and gastrointestinal tissue are highly irregular compared to the pancreatic features that were present in the training data, leading to completely arbitrary labeling of the unrecognized tissue areas. To overcome this, these regions can simply be cropped prior to analysis, as performed for our analyses. Islets comprise a small fraction of the pancreatic tissue area, and were labeled by the model as “other” (i.e. neither normal, ADM, or dysplasia), and therefore introduced only minor errors. In addition, the desmoplastic stroma is a prominent and histologically distinct feature of pancreatic disease that is currently unlearned and labeled as "other" tissue.

Greater limitations arise with the appearance of high-grade neoplasia and adenocarcinoma, both of which can adopt ductal or disorganized structures more closely resembling ADM. It should also be noted that the tool currently labels all non-neoplastic ductal structures as ADM, whether they originate from acinar cells or from ductal cells, and this contributes some error for the quantification ADM of acinar origin. At this stage of the tool’s development, no label for fully developed adenocarcinoma features were used, so lesions that have progressed beyond high grade dysplasia would likely be mislabeled as either ADM or “other”. With future work, it should be possible to train models to identify these additional tissue features and predict them accurately alongside the existing models. The final limitation of the tool is its failure to make accurate predictions in areas of tissue folding or out of focus imaging, but these are obstacles for any image-based measurement

tool (including human annotators) and are avoidable with good technique.

Further work is in progress to reduce error and allow for a broader range of tissue interrogations, including training the tool to recognize a greater diversity of cell types and tissue features such as islets of Langerhans, neural tissue, desmoplastic stroma, adenocarcinoma, and peripheral elements such as lymph nodes or gastrointestinal tissue. The model's quantitative capabilities can also be applied to other disease states that share similar histologic features, such as pancreatitis. Continued development can yield a single comprehensive tool for predicting and labeling all histologic features in pancreatic tissue without the need for complex staining.

Despite the current limitations discussed above, the tool developed here demonstrates clear advantages and superiority to immunostaining for disease quantification in pancreatic pre-cancers. By relying on H&E staining alone, the data acquisition is not only faster and cheaper, but less vulnerable to variable and uneven staining across tissue sections. This consistency and stability of H&E staining eliminates a primary source of error and bias in feature quantification because of manual adjustments needed to threshold immunostained tissues; tissue immunostaining quality varies significantly within single tissue sections and among the many tissues acquired and stained from animal cohorts, typically stained on different days, months, and even years. This tool's exploitation of H&E staining not only enables easy quantitative comparisons between tissues collected and stained across broad time periods, but also enables such comparisons among tissues collected and stained in different laboratories around the world. This unifying aspect will improve collaboration and cross-validation between experiments conducted

by different groups.

Being computer driven, the tool easily quantifies whole pancreatic tissue sections, allowing greater volumes of data acquisitions and avoiding the manual selection of "representative" regions for quantification, which introduces further bias. Furthermore, as an automated, machine-driven measurement tool, potential investigator bias is excluded from the data quantification pipeline. Finally, and importantly, this tool has been demonstrated to identify and segregate key histologic features which immunostaining methods cannot reliably distinguish (i.e. ADM and dysplasias), significantly extending the power of available tissue analytics. This genre of tool will certainly enhance, and conceivably fully replace immunostaining in many animal studies.

2.5 Methods

2.5.1 VISTA Datasets

Murine pancreatic tissues displaying a range of pre-cancerous lesions were isolated from the $P48^{+/Cre}; LSL-KRAS^{G12D}$ mice (KC) mouse pancreatic cancer model. This a widely used genetically engineered mouse model of oncogenic Kras-driven pancreatic adenocarcinoma that closely models the evolution of the human disease, displaying the early hallmarks of ADM, Dysplasia, and desmoplasia, and eventually invasive adenocarcinoma after more than one year of age[45]. Tissue sections from 3 whole pancreases were acquired from KC mice at 5 months for models training. This labeled dataset was split into training

(80%), validation (20% held out from training), and a first testing dataset (20%). Whole pancreas sections from an additional 25 mice were collected at 2 and 5 months of age (n=12, n=13) for IF Spearman correlation testing on a second unseen dataset. Collected pancreases displayed abundant pre-cancerous lesions but were preceding the development of adenocarcinoma. Acute pancreatitis (induced in mice by injection of the pro-inflammatory agent) and normal pancreas sections (n=6, n=3) were also collected for generalizability testing on a third unseen dataset. All pancreas tissue sections were stained with H&E and the second testing set was also stained by immunofluorescence for amylase, labeling normal acini, panK, labeling primarily the oncogenic Kras-transformed epithelial population, and DAPI, labeling all nuclei. These stains were chosen as they are known and well-established markers in the pancreas. Amylase (AMY) is a secretory product of acinar cells, cytokeratins (panK) are well characterized pancreatic ductal lineage markers[115], and DAPI stains cell nuclei which is used as whole tissue area marker. I use AMY, panK and DAPI combination to identify acinar cells from ADM and PanIN tissues. Acinar cells are positive on AMY but negative for panK; ADM tissues are negative for AMY but positive for panK; PanIN tissues are negative for AMY but positive for panK. Normal acini, ADM, and PanINs have nuclei and can be stained with DAPI.

2.5.2 H&E staining and immunofluorescence of Mice Pancreas

The pancreatic tissues were paraffin-embedded, sectioned at 5μm thickness, and stained by standard protocols at the OHSU Histopathology Core. For immunofluorescence staining of amylase and panK, antigen retrieval was

performed using Dako Target Retrieval Solution at pH 9 (Aligent: S236784-2) according to manufacturer's instructions. Specimens were blocked with blocking buffer (1X PBS/5% normal serum/0.3% Triton X-100) for 1 hour at room temperature. The anti-amylase (Santa Cruz: sc-12821) and anti-pan-Cytokeratin (Santa Cruz: sc-15367) primary antibodies were incubated overnight at 4°C, then washed and incubated with secondary antibodies (Invitrogen: A10042 and A32814) for 1.5 hours at room temperature. Slides were covered by coverslips with DAPI's Prolong gold anti-fading agent (Invitrogen: P36931). Fluorescent images of amylase (A), panK (B), and DAPI (C) staining were acquired using a Carl Zeiss Axioscan Z1 slide scanner at a resolution of 0.2 microns/pixel and converted to BigTiff format.

Immunofluorescence images were quantified using ImageJ software. The threshold tool was applied manually to select the amylase-, panK-, or DAPI-positive tissue regions by trained experts. Lymph nodes were manually cropped and excluded.

Despite all data coming from internal sources, steps were taken to better ensure and test the generalizability of models. Each sample of and IF were collected and stained on different days over the course of several month, and samples were taken at different stages of disease progression. Although H&E samples were stained by the same Histopathology Core, it is likely that staining was done by different operators and used different machines. Following model development, generalizability and robustness to H&E staining differences were tested using synthetically altered H&E stains to show model consistency (Figure 2.10). Synthetic HE stains were created by randomly shifting the R, G, and B channels by up to +/- 25% and applying

Gaussian noise. Dice scores were calculated against the unperturbed model predictions. The high mean dice scores support that the model is self-consistent across stains.

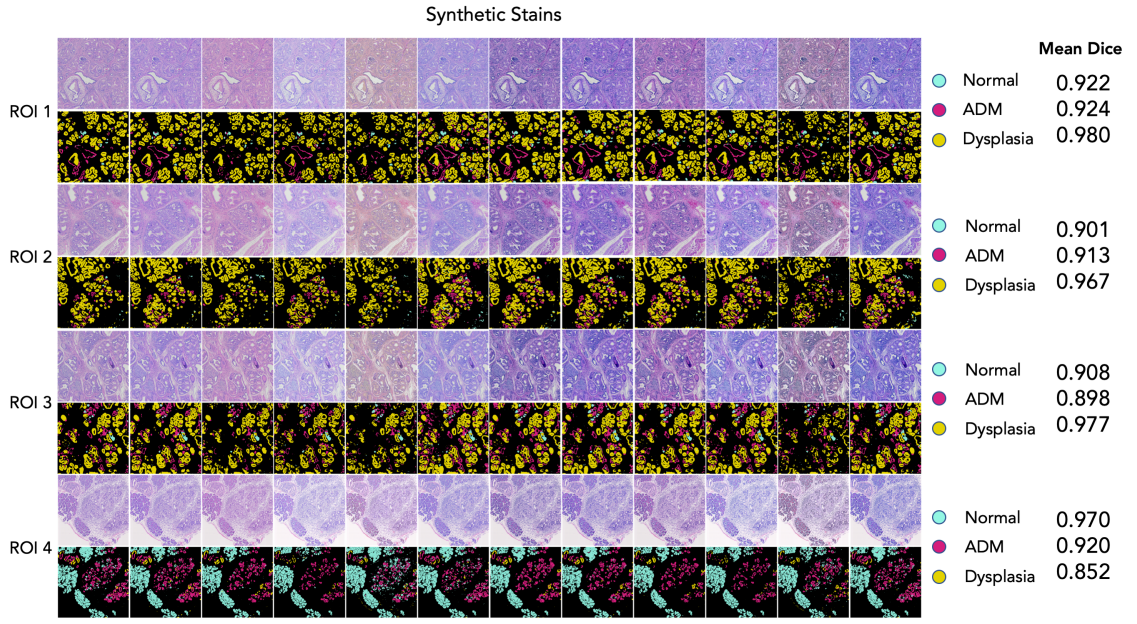


FIGURE 2.10: Synthetic H&E stains were created by randomly shifting the R, G, and B channels by up to +/- 25% and applying Gaussian noise. Synthetically stained images were then passed through the same prediction pipeline to test model robustness across staining qualities. Dice scores were calculated against the unperturbed model predictions. The high mean dice scores support that the model is self-consistent across stains.

2.5.3 Expert annotation of histology features

Annotations for pancreatic tissue features were constructed in Cytomine[82] by three trained experts, and affirmed by a pathologist. These annotations came from 5 regions across 3 images (Figure 2.9) and included at total of 1924 normal acinar, 2582 ADMs, and 1732 Dysplasia (Table 2.3).

Number of Annotations			
	Normal Acinar	ADM	Dysplasia
Image 1	119	1722	1659
Image 2	1342	597	70
Image 3	463	263	3
Total 3	1924	2582	1732
Annotation Area (mm^2)			
Image1	0.05	0.59	1.22
Image 2	0.70	0.17	0.12
Image 3	0.10	0.12	0.02
Total Area	0.85	0.88	1.36

TABLE 2.3: Number of Training Annotations

2.5.4 Training image preparation

In order to make the images more amicable to training for the Deep Learning algorithms, they were trained with intensity normalization to make them appear more consistent with each other. To overcome differential staining across an H&E image, various normalization approaches were applied on intermediate sized (5000x5000 pixel) overlapping crops prior to tiling (512x512 pixel). Background intensities were also ignored from the normalization process to reduce drastic changes on edge regions, isolating only the areas of interest for normalization. Background area was selected by thresholding pixels where all RGB values were greater than 200. The best normalization method was shown to be Reinhard normalization[98] (Table 2.2), so it is used in the implementation of the models.

2.5.5 UNet training

A separate UNet model was trained for each annotated ductal tissue type (normal acinar, ADM, and Dysplasia)[102]. To make each model specific to its respective tissue type, each model's training set was made to incorporate small portions of the other tissue types as negative controls. The training sets were made using 80% of

the total relevant tissue tiles and ∼5-10% of the total of other tissue tiles. Tiles were augmented during training with flips, rotations, and shears to overcome the small dataset size. Training for all three models lasted for 50 epochs, used a batch size of 32 tiles and had a learning rate of 7e-4, implementing the Adam optimizer. Binary cross entropy (Equation 2.1) was used as the loss function during training. Dice Coefficient (Equation 2.2) was used following training to select the best models.

$$BinaryCrossEntropyLoss = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}) + (1 - y_i) \cdot log(1 - \hat{y}_i) \tag{2.1}$$

$$DiceCoefficient = \frac{2(X \cap Y)}{|X| + |Y|} \tag{2.2}$$

where X is the ground truth segmentation mask and Y is the predicted segmentation mask.

2.5.6 Model integration

As a standard, models produced through Deep Learning packages will call anything with a prediction value >= 0.5 as positive and anything < 0.5 as negative. This threshold, however, might not be the ideal and can be subject to optimization and tuning. Within the training and validation datasets, it was noticed that the standard thresholds led to pixel level false positive noise and predictions that bleed into surrounding ductal lumen. To make the models more accurate, thresholds were chosen based on the Receiver Operating Characteristic (ROC) curves (Figure 2.11) – sensitivity and specificity, and were manually adjusted to reduce

the observed errors qualitatively. This step would help to ensure that the models would better generalize to the testing dataset with minimal noise, taking only predictions the model was most confident in. Within the testing a validation set, the following thresholds were chosen for each model respectively, and the chosen thresholds were carried forward to be used in subsequent testing: Normal Acinar Threshold = 0.3, ADM Threshold = 0.5, and Dysplasia Threshold = 0.7.

After manual parameter tuning, the determined thresholds remain within a reasonable range, as observed by the ROC curves. Once each model made its prediction for a given tissue, the background white pixels were again removed from prediction by ignoring all pixels where all RGB values were greater than 200. Total tissue (DAPI positive) region was also calculated by finding all pixels where RGB values were lower than 200. To combine all four tissue masks, normal acinar predictions override metaplasia and dysplasia predictions; metaplasia predictions override dysplasia predictions; normal acinar, metaplasia, and dysplasia predictions all override DAPI predictions.

2.5.7 VISTA validation and testing

Because no foreign tissue was used for negative controls during training (primarily lymph nodes and GI tissue), regions of testing images containing these tissues had to be cropped out prior to testing and analysis. Testing and analysis were performed through a similar pipeline as training, incorporating intermediate crop normalization and tile level prediction. These overlapping tiles were stitched back into a full image and an average was taken to get pixel level predictions for

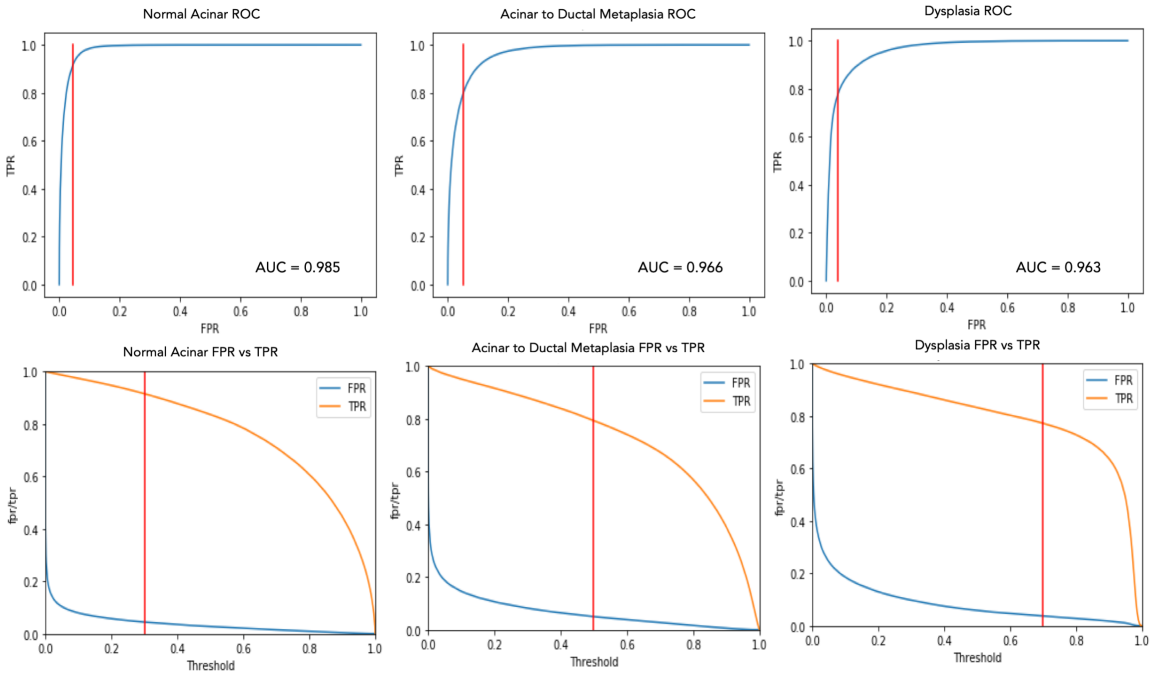


FIGURE 2.11: ROC curves were created for each model by testing the error rate at 0.1 increments. The red line represents the chosen threshold's corresponding false positive/true positive rates. The high AUC values indicate that the models are highly accurate. The fact that the chosen threshold falls well within the elbow of the ROC curve indicates that despite manual adjustment, the chosen thresholds hold their high predictive value.

each model. In validation, models were compared to annotations from the held-out dataset of labeled images. In testing, model predictions were compared to three unseen datasets: the first comprised of labeled tiles, the second comprised of immunostained serial sections that were thresholded by an expert, and a third comprised of normal and pancreatitis whole tissue sections. To compare with immunostaining, ADM and dysplasia predictions were combined to make a general panK prediction mask. Predictions were then paired with their respective serial section and correlated to determine model accuracy. Correlation was chosen as the metric for this test over Dice or sensitivity because serial sections have a 5 µm offset which causes the H&E used for predictions and the IF used for ground truth to

be spatially unaligned. Although correlation of abundances remains high between serial sections[21], the errors in alignment have strong negative biases on metrics like Dice even after attempts at registration. Using IF as ground truth also adds biological credibility to the metrics while annotated ground truths were found to be prone to annotator error.

Like what is done with other virtual staining methods that have been deployed on tissue sections[100, 101, 15] I also evaluated our predictions against IF staining was also done with structural similarity:[143]

$$SSIM(x,y) = \frac{(2\mu_x\mu_y)(2\sigma_{xy} + (k_2L)^2)}{(\mu_x^2 + \mu_y^2 + (k_2L)^2)(\sigma_x^2 + \sigma_y^2 + (k_2L)^2)} \tag{2.3}$$

where x and y are input images, μ_x and μ_y are the mean intensities, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance, k_1 and k_2 are constants, and L is the dynamic range. SSIM was calculated using the scikit-image library with all default parameters:[139] sliding window size = 7 pixels; k_1 = 0.01; k_2 = 0.03; and data range estimated from images. The SSIM between two images is calculated over pixel neighborhoods in the images and provides a more coherent measure of image similarity than pixel-wise measures. I chose SSIM as the metric for comparison because it would be more robust than Dice against differences between the serial H&E and IF sections, and small Gaussian blurs were applied to account for tissue differences at the pixel level. The range of SSIM values extends from -1 to +1, and only equals 1 if the two images are identical. Values close to one are indicative of good reconstruction and strong model performance. Four ROIs from two whole slide test sections that had high correspondence between H&E and IF were used

for analysis.

The amylase, panK and DAPI area were measured in pixels, and the percentage of positive areas were calculated as a percent of the total all measured cellular regions.

2.5.8 Statistical analysis

Since datasets were continuous, independent, and had no tied values, after checking the assumption and conditions were met, and since the datasets were small, non-Gaussian, and contained outliers, the non-parametric Mann-Whitney U test was used to access statistical differences in means. Since datasets were small and had outliers, the correlation tests for all models were conducted using Spearman correlation.

2.5.9 Animal models

This work was performed in accordance with Institutional Animal Use and Care Committee (IACUC) guidelines of the Oregon Health and Science University (OHSU). All work involving mice received approval by the IACUC at OHSU. The KC mice were all backcrossed at least 5 generations into the C57Bl6/J background. Acute pancreatitis was induced in 6-week old C57Bl6/J mice by intraperitoneal injection of 50 µg caerulein (Sigma:C9026) per kg body weight, with a total of 7 consecutive treatments at 1hour intervals. Pancreatic tissues were harvested 3 days following caerulein treatment. Caerulein was dissolved in PBS at a concentration of 10 µg/mL.

Chapter 3

Extracting novel, biologically relevant features from images

Men honor what lies within the sphere of their knowledge, but do not realize how dependent they are on what lies beyond it.

Readings from Chuang Tzu

3.1 Abstract

Image-based cell phenotyping relies on quantitative measurements as encoded representations of cells; however, defining suitable representations that capture complex imaging features is challenged by the lack of robust methods to segment cells, identify subcellular compartments, and extract relevant features. Variational autoencoder (VAE) approaches produce encouraging results by mapping an image to a representative descriptor, and outperform classical hand-crafted features for

morphology, intensity, and texture at differentiating data. Although VAEs show promising results for capturing morphological and organizational features in tissue, single cell image analyses based on VAEs often fail to identify biologically informative features due to uninformative technical variation. Herein, I propose a multi-encoder VAE (ME-VAE) in single cell image analysis using transformed images as a self-supervised signal to extract transform-invariant biologically meaningful features, including emergent features not obvious from prior knowledge. I show that the proposed architecture improves analysis by making distinct cell populations more separable compared to traditional and recent extensions of VAE architectures and intensity measurements by enhancing phenotypic differences between cells and by improving correlations to other analytic modalities.

3.2 Introduction

Understanding cellular changes and phenotypic pathways at the single cell level is becoming increasingly important because it creates a comprehensive understanding of cell state and cell-to-cell heterogeneity. Multiple analytical tools are available to extract, normalize, and evaluate single cell RNA sequencing (scRNAseq) data[130, 107, 11]. Until recently, analyzing single cell imaging data in a similar fashion was limited to extracting mean intensity profiles, predefined shape, textural and morphological features, and images stained with only a few markers. Emerging multiplexed imaging technologies such as cyclic immunofluorescence (CyCIF)[68], multiplexed immunohistochemistry[117], CO-Detection by indEXing (CODEX)[25], and Multiplexed ion beam imaging[5] create images comprised of

a large number of markers, expanding the depth of information. Robust analytical methods for high-dimensional multiplexed imaging data, however, are still needed. One limitation with analyzing highly multiplexed single cell images is the ability to extract biologically meaningful information on staining localization patterns that indicate divergent cell states. Single cell imaging data has morphospatial information not captured using simple mean intensity information, with successful quantification of these features potentially leading to improved analysis and understanding[145].

The classical approach for image feature extraction is manually creating a list of desired features and predefined metrics to quantify them. This is biased toward known and easily measured features and can miss subtle but important features. More robust image feature extraction has been employed using deep learning architectures such as the Variational Autoencoder (VAE)[58] in other domains where feature representation can be automatically generated without supervising information or prior knowledge.

However, the problem with VAE feature extraction in single cell imaging is that there are typically unimportant or uninformative technical features driving differences between biologically similar images such as image transformation features like rotation, offset within the image, affine/skew, and stretching. These features are extracted and entangled with the other features from the VAE leading to the downstream analysis being skewed in undesired ways[35]. Despite having the same underlying information, the common uninformative features in transformed images distract deep learning architectures so that they ignore most of the biologically relevant features[48, 8, 155, 84, 42]. This holds true in single cell images,

where VAEs frequently ignore biologically meaningful features and focus on recreating the transformational features which have a high variance across the dataset. When these features are known controllable transformations, they can be used for a self-supervised signal to extract invariant features with respect to a set of transformations during model training. This means that the model does not need a ground truth annotation of what is being learn becasue it can supervise its own learning progress using the results of different but biologically similar inputs. Untailored deep learning architectures are unable to overcome these uninformative features unless some modification is made to either their architecture or objective functions[48, 8, 155, 84, 42]. Many recent works propose changing autoencoder architectures to coupled networks or using multiple latent dimensions to overcome this without the need for biased hyperparameter tuning and data normalization[35, 146, 41, 41, 34]. Similar methodologies have also been explored that seek to correct transformative features with coupled networks, direct latent space modifications, novel layer architectures, and training networks with combinations of corrected and uncorrected image data[48, 8, 155, 84, 42, 106, 93]. Most of these corrected architectures, however, only target to one specified feature and can't generalize to other features without further modification. Two examples of recent architectures that use modifications to the objective function are the β-VAE[44] and the invariant C-VAE[86], which use their loss functions to pressure the model such that it will prioritize a more learning more quantified features that are more interpretable, balanced, and/or invariable to specific features.

Here I propose a novel method for single cell image feature extraction that removes specified uninformative features by making them uniform and invariant across the

reconstructions, using modified pairs of transformed input and output images by self-supervised transformation, and utilizing multiple encoding blocks. Using this multi-encoder VAE (ME-VAE) to control for multiple transformational features, I highlight its ability to extract biologically meaningful and transform-invariant single cell information and better separate biologically distinct cell populations without the need for biased manually selected feature sets.

3.3 Results

3.3.1 Controlling for uninformative features

When a transformational feature varies across a single cell imaging dataset, standard VAEs extract only the dominant component to reconstruction. When rotation varies from image to image, reconstructions along the principal component walk[109] only constitute the angle of the cell and downstream analysis is heavily skewed by this extracted component (Figure 3.1a). In another dataset where polar orientation is the dominant feature, I observe the same behavior (Figure 3.1b); VAEs only extract the dominant uninformative features, ignoring subtle but informative features necessary for detailed reconstruction.

In order to overcome model hypersensitivity to dominant uninformative features, several architectures were proposed and tested to learn the latent space while attempting to ignore uninformative features (Figure 3.1c-g). A standard VAE without control for uninformative features was used as baseline and shows a high correlation between the embedded components and the respective feature metrics

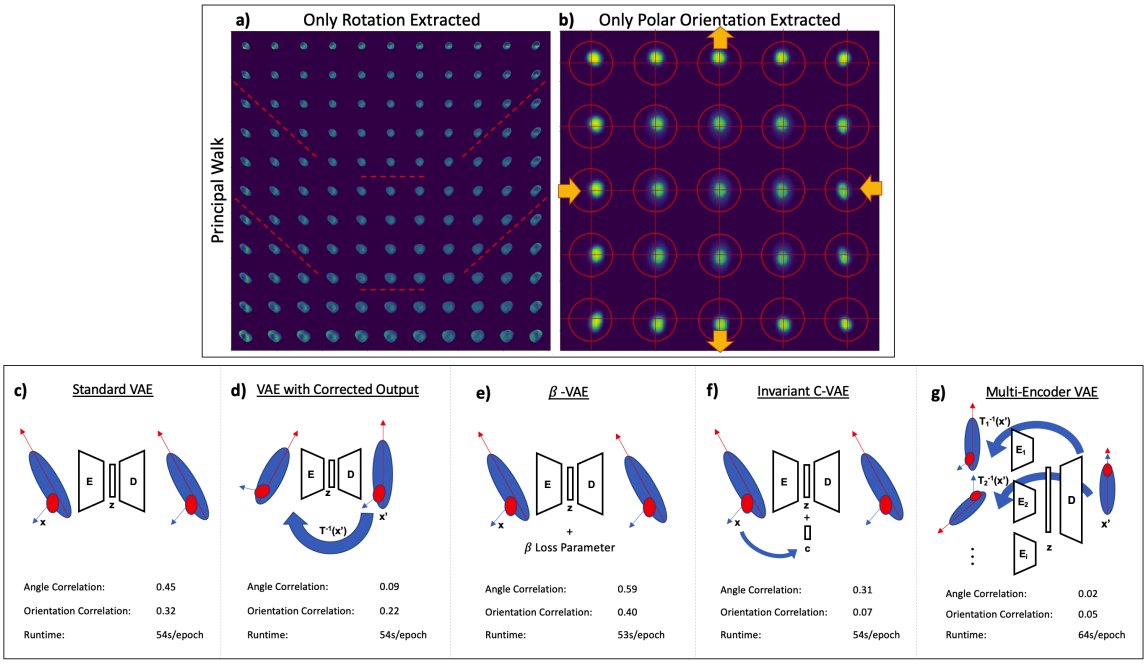


FIGURE 3.1: VAE analysis of two datasets are shown, each governed by a single biologically uninformative feature a) rotation and b) polar orientation. Principal walk reconstructions[109] show the VAEs' governing features across the latent space through a range of image reconstructions. To correct this model hypersensitivity several architectures were tested: c) standard VAE with matched raw images; d) VAE with paired randomly transformed input and controlled output images; e) β-VAE which operates similarly to the Standard VAE, but utilizes a λ hyperparameter in the loss function to encourage an independent latent space; f) an invariant conditional VAE that injects the values of the uninformative features into the decoder such that they are not embedded in the latent space; g) the proposed multi-encoder VAE: VAE with corrections for multiple features (rotation, polar orientation, size, shape, etc.) using parallel encoder models, a shared latent space, and a single decoder model. In c)-e), a correlation between the embedding components and the respective feature (angle and orientation) is measured to quantify how effectively the model removes uninformative features. PBS and TGFβ+EGF cell populations with single channels were used in this analysis (n = 15,898 single cell images).

(Figure 3.1c and Figure 3.2a). When a single factor is controlled (e.g. rotation), it becomes uncorrelated to all VAE encodings, and even the max correlated component in the latent space is insignificant (Figure 3.1d and Figure 3.2b). Controlling for one feature does not significantly impact the other dominate transformation features (i.e. polar orientation). With a double transformed output correcting two

features simultaneously, there is a decorrelation of both dominant features (Figure 3.2c), but the reconstructed images are poor (Figure 3.2g) reflecting the model's failure to learn relevant feature embeddings. The VAE with transformed output is shown to work on simple transforms such as rotation, but pairs of complex transformations like rotation combined with polar orientation prove too difficult. Both the β-VAE[44] and invariant C-VAE[86] also show strong correlations between the uninformative features I wanted to ignore and the latent space (Figure 3.1e/f and Figure 3.2d/e). Finally, when both uninformative features are controlled for using the proposed ME-VAE with transformed image pairs, there is a decorrelation in both uninformative features, indicating that the VAE reconstructions learned to overcome them and focus on underlying features that better separate cell populations (Figure 3.1g and Figure 3.2f). Unlike with the corrected output VAE, the ME-VAE produced coherent reconstructions (Figure 3.2h), as can be seen by the qualitatively well-defined and more realistic images as opposed to the messy images with no clear biological pattern retained. Moreover, the ME-VAE is better able to generalize to new datasets and is scalable since it controls many uninformative features together in parallel by using a multi-encoder network where any number of encoders can be added, and each encoder learns a single transformation. Finally, when training on the same dataset of 15,898 single channel images, all comparison architectures took a similar amount of time to train ranging between 53 and 54 seconds per epoch on average. The proposed architecture only took a few seconds longer, averaging 64 seconds per epoch, indicating that the increased performance and reduction in uninformative features does not come with a significant increase in computation time.

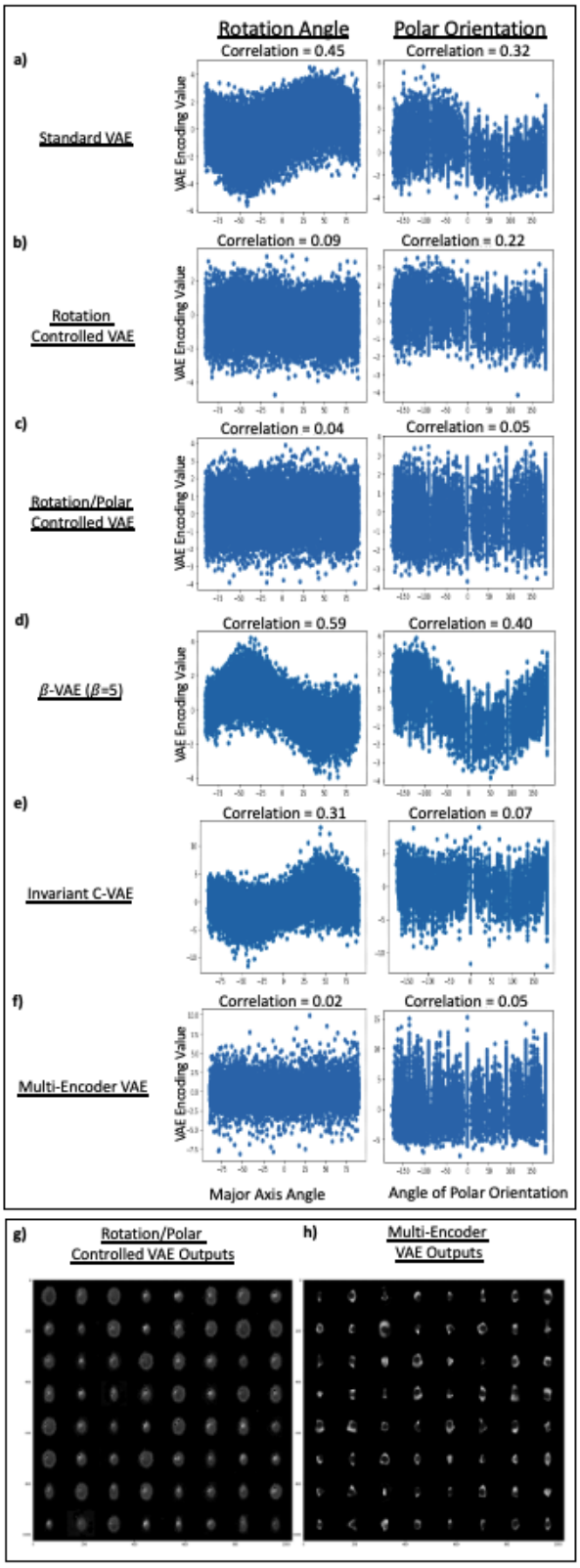


FIGURE 3.2: Encoding spaces for each VAE method were analyzed for correlation with uninformative features. Scatter plot and correlation is shown for the latent space component that had the highest correlation to given the metric. Correlations for all methods utilized a sample size of n = 15,898 single cell images a) Standard VAE used as baseline to show high correlation between encoded features and undesired features. b) Output corrected transform invariant VAE controlling for rotation only. c) Output corrected transform invariant VAE controlling for rotation and polar orientation. d) β-VAE implementing λ hyperparameter in loss function. e) Invariant C-VAE using quantified values of uninformative features injected into decoder. f) Proposed multi-encoder VAE correcting for both rotation and polar orientation. g) Failed reconstruction examples from the transform invariant VAE correcting for both rotation and polar orientation. h) Successful reconstruction examples from the ME-VAE correcting for rotation and polar orientation.

3.3.2 Improving biological interpretation on single channel images

To evaluate the models' abilities to improve downstream usefulness and biological relevance, I analyzed a dataset (subsection 3.5.1) of single cell CyCIF images from MCF10A non-malignant breast epithelium cell line. The full dataset I analyzed is comprised of 6 ligand treated cell populations and is stained with 23 biomarkers. Here, I restricted our analysis to PBS (control) and TGFβ+EGF population and considered only the Epidermal Growth Factor Receptor (EGFR) channel. These were chosen because they have similar distributions of cell size and mean whole cell EGFR intensity following cell level normalization, making them difficult to naively separate with classical cellular features (Figure 3.3a), but qualitatively show phenotypic differences such as compartment localization and stain texture. Within this dataset I show that the ME-VAE better separates PBS and TGFβ+EGF treated cell populations compared to the standard VAE.

As can be observed in Figure 3.3b, the standard VAE is incapable of separating the two cell populations, creating a mix of the labeled cell populations in k-means cluster space (number of clusters = 2) and UMAP embedding space. The cells within UMAP regions also have an arbitrary range of phenotypes; the only observed patterns are of uninformative features such as rotation, polar orientation, and size. Classically extracted features show similar results to the standard VAE (Figure 3.3c) where uninformative and non-biological features govern the clustering and UMAP distribution. Despite the fact that orientation was not included in the set of extracted properties, the rotation angle is still captured because the

FIGURE 3.3: Cells are compared using initial naïve metrics such as mean EGFR intensity and cell size to show the difficulty separating the cell populations. b-f) The model architectures are quantitatively evaluated using cluster purity (k-means with number of clusters = 2). The sample size for all comparison methods and metrics is n =15,898 single cell images. Qualitative comparison is made using visual separation of two labeled cell populations in UMAP embedding space and visual analysis of cells from UMAP regions to identify biologically distinct factors. Rotation angle of cells are shown in UMAP embedding to show the influence of unimportant features on downstream analysis. g) The same population of cells are compared using radial slope (a metric inferred from visual analyzing the regional cell images in c).

same information is available through a combination of important features such as eccentricity, extent, moments, and inertia which were extracted. The β-VAE architecture[44] does not show significant improvement from the standard VAE either (Figure 3.3d). Moreover, the λ hyperparameter is known to be difficult to tune which can lead to large variations in both reconstruction quality and clusterability (Figure 3.4). The invariant C-VAE adapted from Moyer et al.[86] does see an improvement in clustering compared to the standard VAE (Figure 3.3e), but despite having the uninformative values injected into the model, it is unable to keep them from being encoded in the latent space, resulting in UMAP embeddings dependent of uninformative features. Many of the recent extensions of the VAE that seek to improve the interpretability of the latent space simply modify the loss function used during training to encourage a result instead of forcing it (subsection 3.5.2). Unlike these previous attempts, the ME-VAE changes the actual deep learning architecture by adding multiple encoding blocks each for the purpose of removing a specific feature, which I observe to has an increased performance.

By comparison to all other attempted methods, the ME-VAE has a dramatic increase in k-means cluster purity and normalized mutual information (NMI) and shows a clear separation of labeled cell populations in UMAP (Figure 3.3f), indicating improved clusterability and separability. Regional cell images within the multi-encoder's UMAP space show distinct phenotypic differences that separate the cell populations with biologically relevant features (stain localization, intensity, and subcellular pattern). In PBS dominant regions, EGFR stain is most heavily concentrated uniformly along the cellular membrane, while TGFβ+EGF regions show

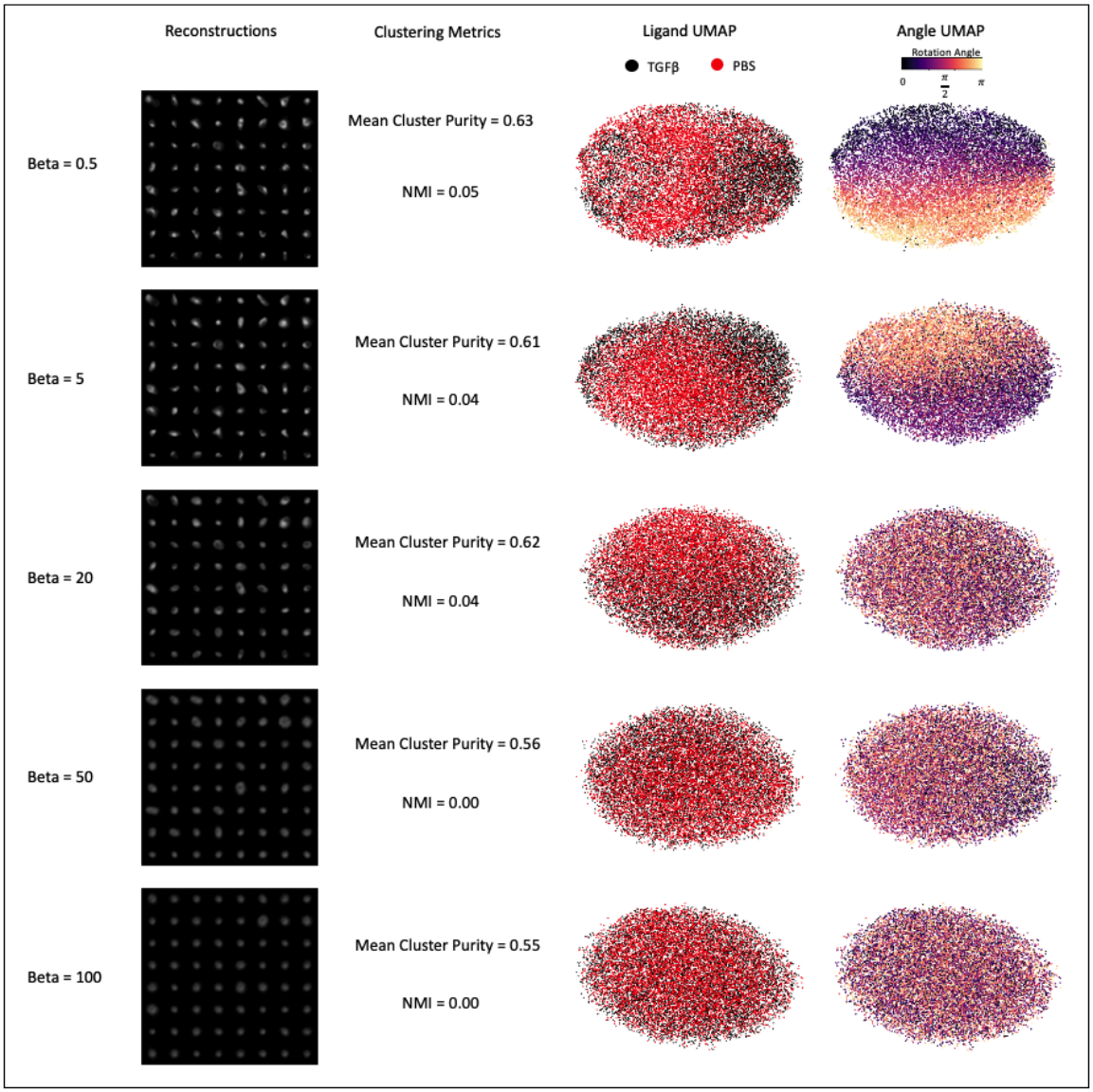


FIGURE 3.4: Left panels are shown 25 randomly sampled image reconstructions across varying values of β, followed by their quantified clustering metrics generated using k-means with number of clusters = 2 and sample size of n = 15,898 single cell images, and on the right are the models' projections into UMAP, colored by ligand population and rotation angle.

a cloudy diffuse concentration of EGFR stain throughout the cell with the heaviest concentration of stain localizing to one side of the nuclear membrane. These

differences illustrate a clear difference in cellular regulation and compartmentalization of the EGFR protein induced by the TGFβ+EGF ligand combination. Based on observations from the multi-encoder output in Figure 3.3f, I inferred the metric of radial slope would similarly separate the two populations (subsection 3.5.3, Figure 3.5). I observe a larger (more positive) radial slope in the PBS population on average, indicating that the distribution of stain increases radially toward the membrane, and by comparison the radial slope of the TGFβ+EGF population has a smaller (more negative) radial slope than the PBS, indicating that the stain distribution is located primarily toward the center of the cell and decreases radially toward the membrane. Using this metric, there is improved separation and cluster purity and NMI compared to the selected naïve metrics (Figure 3.3a and Figure 3.3g). What's more is that the concentration of EGFR in the TGFβ+EGF population is located just outside the nucleus, and therefore would not be successfully separated simply be isolating the mean intensity of the nuclear region. The clustering metrics from radial slope, however, are still lower than the full ME-VAE cluster purity, indicating more features beyond the radial slope are being extracted from ME-VAE.

3.3.3 Use case with a large complex dataset

Models were next trained on the expanded dataset (five ligands and PBS control) and 23 channel CyCIF images (subsection 3.5.1 and Table 3.1). Like before, the ME-VAE was trained to control for rotation, polar orientation, and cell size/shape, and the standard VAE performed similarly to the previous experiment, encoding cells based primarily on the dominant features such as size and rotation while

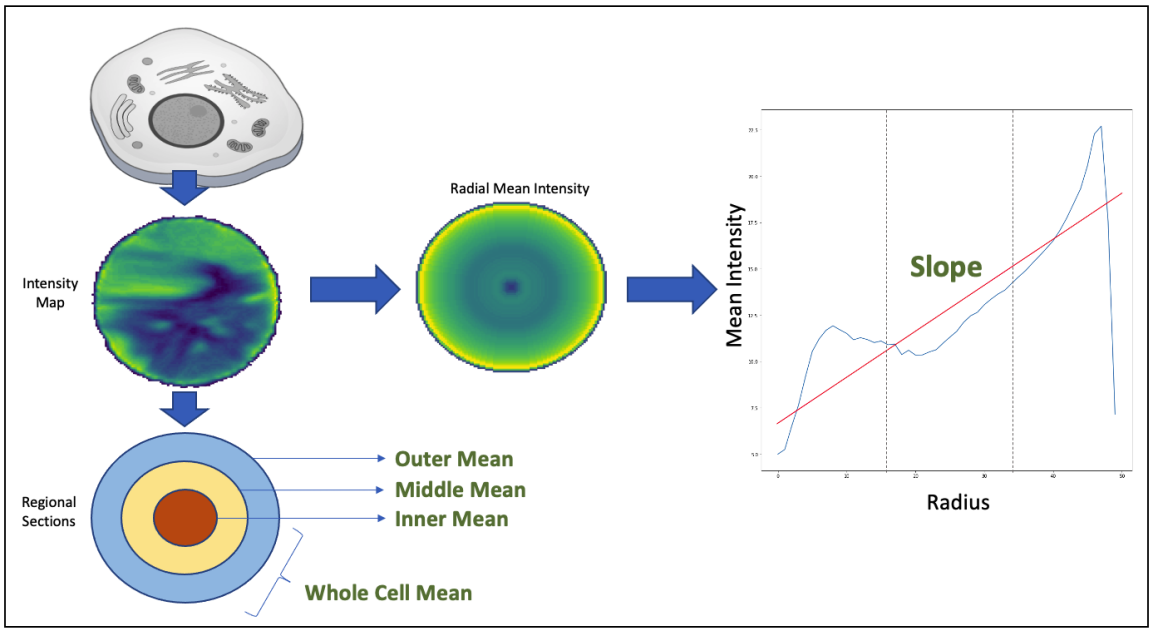


FIGURE 3.5: Cell intensity maps were circularized to allow easy compartmentalization. The inner, middle and outer mean intensities were extracted by dividing the cell into thirds radially. The mean intensity of the whole cell was also taken. The radial mean intensity map was created by taking the average intensity for each radius across the circularized cell. The slope of the radial mean intensity map was then taken to create a single metric for stain distribution.

largely ignoring complex staining information (Figure 3.6a,b – left). Although visually there is some preferential localization in UMAP (OSM left side, TGFβ+EGF right side), it is clear that the populations are thoroughly mixed with poor separability. The intensity profiles show that size has a strong impact on this left/right embedding (Figure 3.6b left). Most stains show little or no consistency within the embedding space, with the exception of DAPI and Ki67. These stains, however, show the same left/right distribution as size, indicating the nuclear intensity distributions are simply a result of cell size, since the whole cell mean intensity of a nuclear marker will decrease with larger cells and increase with smaller cells.

Channel	Marker
1	DAPI
2	STAT1 (p-S727)
3	Vimentin
4	Cytokeratin 7
5	ki67
6	S6
7	LC3A/B
8	NFkB (p65)
9	p21 (Waf1/Cip1)
10	Catenin (Beta)
11	S6 (p-S235/S236)
12	PDL1
13	E-cadherin
14	STAT1 (alpha-isoform)
15	HES1
16	EGFR
17	NDG1 (p-T346)
18	STAT3
19	S6 (p-S240/244)
20	MET
21	Cytokeratin 18
22	Cyclin D1
23	c-Jun

TABLE 3.1: CyCIF Marker Panel

Despite the increased complexity of the multi-channel CyCIF images and diversity of the dataset which could overload a simple architecture, the ME-VAE shows good separation of the labeled cell populations (Figure 3.6a – right). I also observe subcluster formation for HGF, BMP2+EGF, and TGFβ+EGF. By analyzing intensity profiles and regional cell images of these populations, one can see differences in expression (Figure 3.6b right and Figure 3.7b – right). The UMAP intensity profiles show clear stain intensity patterns indicating that the ME-VAE encoding space contains relevant biological information. Size does show some distribution in the UMAP, but the effect is largely dulled in comparison to the standard VAE.

Here I discuss some of the most noticeable drivers of separation between cell populations in the MCF10A dataset. PBS shows a marked decrease in Ki67 expression compared to other ligands, consistent with a relative decrease in proliferation. The

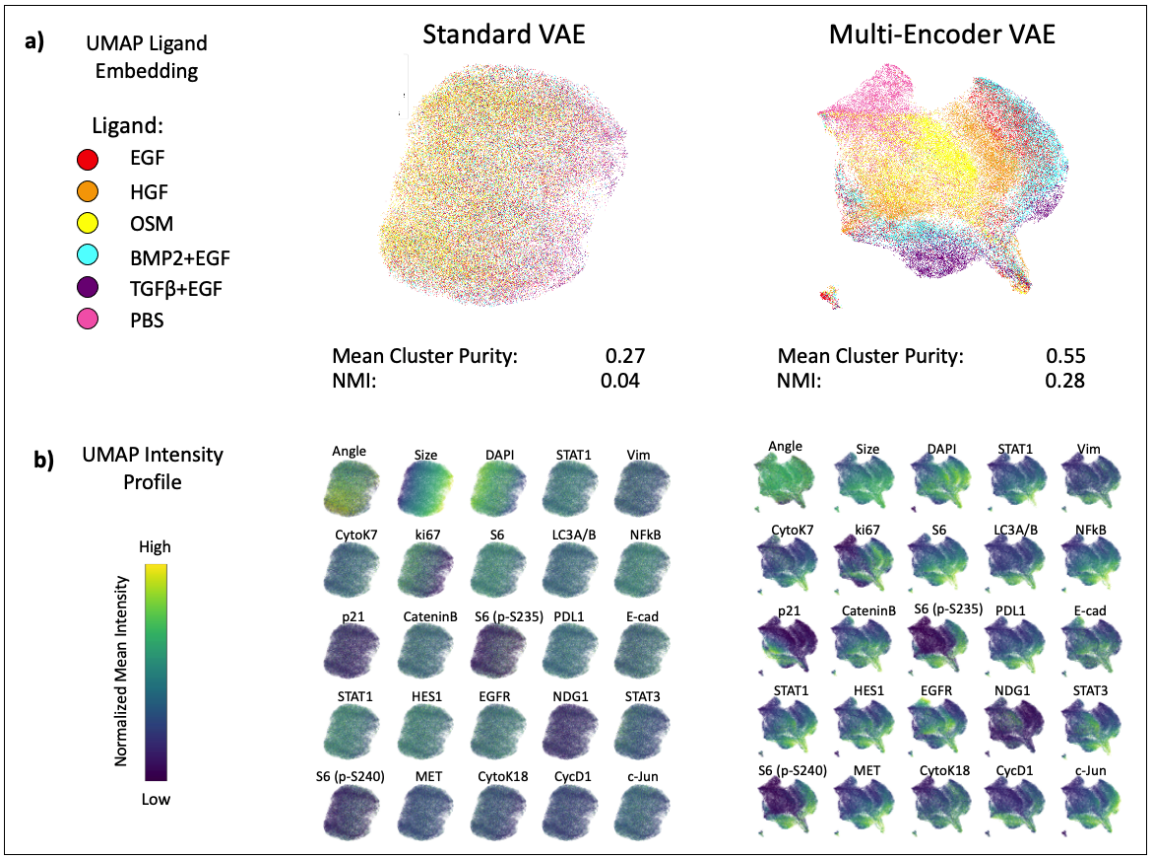


FIGURE 3.6: a) UMAP embeddings for respective VAE encodings, allowing for qualitative visual evaluation of ligand separability. Mean cluster purity and NMI was calculated to quantitatively compare methods (k-means number of clusters = 6). Total sample size is n=73,134 single cell images. b) Distribution of stain features across UMAP space, colored by intensity.

TGFβ+EGF populations show an increase in S6 expression, indicating an increase in cell growth. This is observed visually with regional cell images (Figure 3.7b right); however, it's worth noting that high S6 expression is seen in both large and small cells treated with TGFβ+EGF. In EGF and BMP2+EGF treated populations, decreased expression of membrane adhesion proteins such E-cadherin and β-Catenin is observed. This decrease presents visually as dim stain, but the marker is still localized to the membrane rather than missing or diffuse throughout the cell.

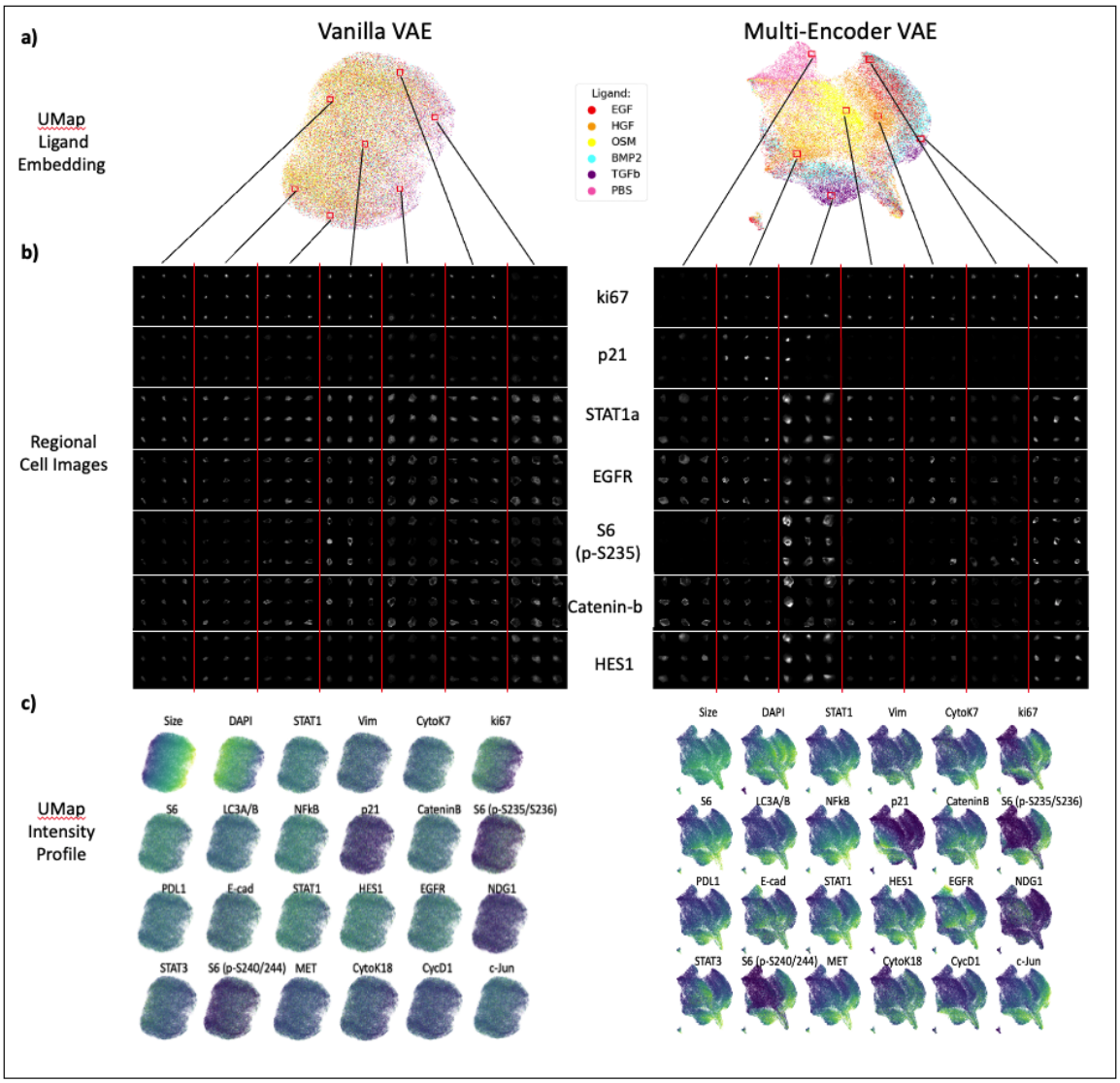


FIGURE 3.7: a) UMAP embeddings for respective VAE encodings, allowing for qualitative visual evaluation of ligand separability. b) Regional cell images were sampled from locations throughout UMAP space to highlight the differences in expression pattern. Stains shown were selected based on a combination of being correlated to important VAE features and hand-selection for known variance.

In both TGFβ+EGF- and PBS-treated cells, there is increased concentration of HES1 localized primarily to the nucleus, while in other populations the distribution is

uniform throughout the cell. In the case of TGFβ+EGF, this localization is accompanied by increased intensity (Figure 3.7b right), but PBS intensity is more similar to the other ligand-treated populations. Similarly, Stat1a is primarily located in the nucleus for TGFβ+EGF-, BMP2+EGF-, and OSM-treated populations, but shows decentralized staining in cell images for other ligand populations. This is important because both HES1 and Stat1a are functional in the nucleus (Stat1 particularly as it translocates into the nucleus as part of its functional pathway) with limited activity in the cytosol[87, 55]. Another observation is that p21 uniquely separates subpopulations in TGFβ+EGF-, HGF-, and BMP2+EGF- treated cells, indicating that there are subsets of the population that are undergoing growth arrest due to inhibition of cell cycle progression via p21 regulation.

These results show that the ME-VAE captures relevant biological information and separates cell populations, highlighting important features without significant interference from the controlled uninformative features. Furthermore, the ME-VAE can capture emergent biologically relevant imaging features not obvious without prior knowledge. By contrast, little to no biologically relevant information is obtained from the standard VAE.

3.3.4 Correlation of reverse phase protein arrays pathway activity and CyCIF using ME-VAE features

To validate that ME-VAE yields biologically more meaningful representations, I correlate VAE features with respect to Reverse Phase Protein Arrays (RPPA) pathway activity. By reordering VAE features using hierachical clustering to form a

feature spectra, I extract broad patterns and reduce the dimensionality of the feature set. The standard VAE shows very poor self-correlation with only a handful of feature clusters showing strong correlation (Figure 3.8a top). Comparatively, I observe a clear pattern of self-correlations between ME-VAE features, indicating the model successfully extracts distinct yet different expression patterns (Figure 3.9a top). I identify ten representative clusters from the ME-VAE latent space that illustrate different expression patterns, which are explored using representative images (Figure 3.9a bottom). Representative cell images are chosen by selecting the cell for each feature set that has a high mean expression of all features in that respective aggregated feature set. Between clusters 0 and 1, there is a difference in the ratio of nuclear size and cell size. Cluster 1 encodes for larger nuclei than cluster 0 (this pattern is reaffirmed in Figure 3.9b where cluster 1 correlates to DNA pathways and nuclear stains while cluster 0 does not). Cluster 4 is a highly varied cluster but contains large cells with more diffuse intensity patterns. From these aggregated features, one can observe that the ME-VAE architecture extracts a combination of intensity and morpho-spatial profiles with at least 10 clear axes of variation. Using these aggregated features, one can analyze and interpret biological meaning with fewer spurious correlations than comparing many to many.

A growing method for single cell analysis is to integrate multiple modalities. Multi-modal integration helps validate where the two modalities overlap, expands the dataset with mutually exclusive or orthogonal features, and allows for cross-wise mapping of features. This form of integration is also important because biology operates across scales, compartments, and data types (such as genetic, transcriptomic, and proteomic), and across time. Being able to encode

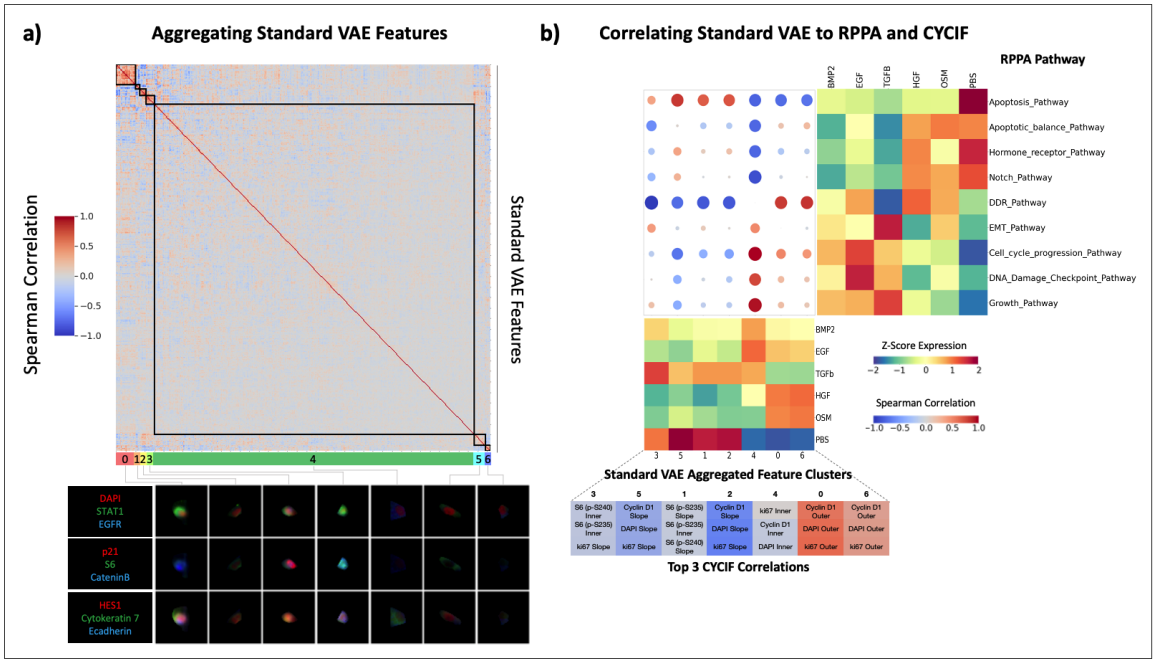


FIGURE 3.8: a) Using the single cell observations as features, correlations are drawn between pairs of standard VAE features. These features are then hierarchically clustered to observe patterns and reduce VAE features to aggregated feature sets. Cell images were assigned aggregated feature scores using the mean expression of each feature in a cluster. Shown are representative cells that are highly expressing for each respective cluster.
b) Correlation matrix between RPPA pathway activity scores and standard VAE aggregated features. Samples from the two modalities were paired by their ligand treatments, resulting in a sample size of n=6 biologically independent ligand treated cell populations. RPPA pathways and VAE features were hierarchically clustered to show prominent patterns in correlation. Standard VAE aggregated features were also correlated to several metrics of CyCIF expression (mean inner, mean middle, whole cell mean, and radial slope) for all 23 stains. This CyCIF correlation was done using the full dataset of single cell images (sample size n=73,134 single cell images)). The table of CyCIF correlations shows the top three correlations for each ME-VAE aggregated feature. Aggregated feature 4 shows high correlations to almost all RPPA pathways (3rd column from the right), and the DNA death/repair and apoptosis pathways also has high correlations to almost all aggregated features (1st and 5th rows).

the full scope of information is necessary for understanding the full picture of biological information. The validation of additional modalities is especially important for VAE-based single cell image analysis because it frames inherently obscure encoding features in a biological context and validates that the extracted

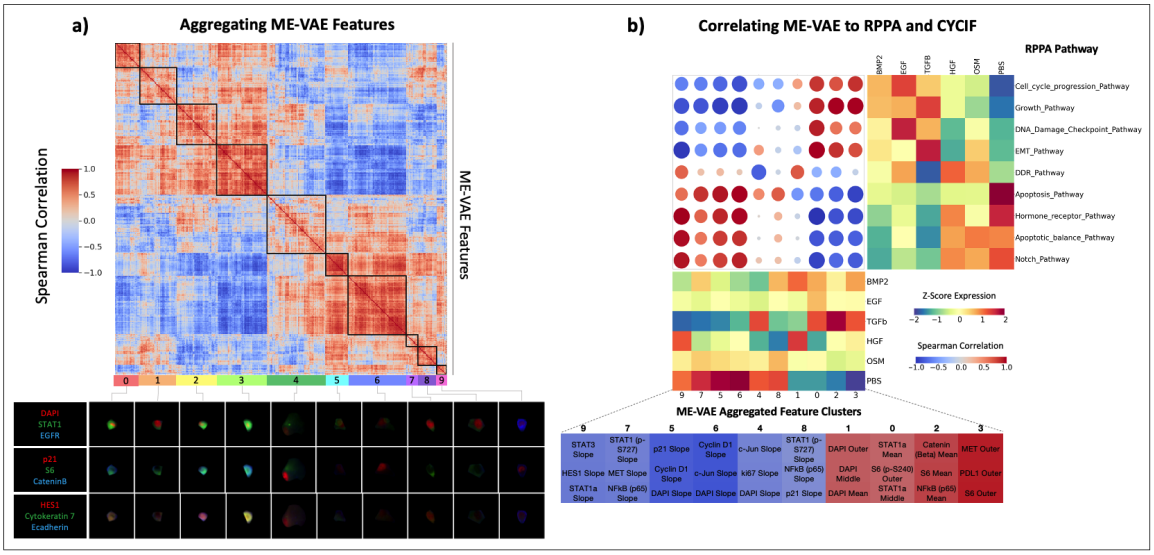


FIGURE 3.9: a) Using the single cell observations as features, correlations are drawn between pairs of ME-VAE features. These features are then hierarchically clustered to observe patterns and reduce VAE features to aggregated feature sets. Cell images were assigned aggregated feature scores using the mean expression of each feature in a cluster. Shown are representative cells that are highly expressing for each respective cluster.
b) Correlation matrix between RPPA pathway activity scores and ME-VAE aggregated features. Samples from the two modalities were paired by their ligand treatments, resulting in a sample size of n=6 biologically independent ligand treated cell populations. RPPA pathways and VAE features were hierarchically clustered to show prominent patterns in correlation. ME-VAE aggregated features were also correlated to several metrics of CyCIF expression (mean inner, mean middle, whole cell mean, and radial slope) for all 23 stains. This CyCIF correlation was done using the full dataset of single cell images (sample size n=73,134 single cell images). The table of CyCIF correlations shows the top three correlations for each ME-VAE aggregated feature.

features are coherent. The increased feature range of ME-VAE allows for cross-wise mapping and integration of complex CyCIF image features and other modalities (e.g. RPPA).

When correlating the 7 aggregated standard VAE features with RPPA pathway activity, one can notice two distinct issues. First, there is a single aggregated feature that shows significant correlations shows correlates to nearly every RPPA pathway activity profile (Figure 3.8b). Second, there is a single RPPA pathway that correlates

to nearly every standard VAE aggregated feature. When correlating standard VAE aggregated features to the extracted CyCIF metrics (subsection 3.5.3, Figure 3.5), the Spearman correlations are small despite the increased sample size of n=73,134 single cell images (Figure 3.8b), with the largest correlations being restricted to nuclear markers such as CyclinD1, DAPI and Ki67. As mentioned above this is likely an artifact of encoding for size since nuclear expressions can be a function of cell size. By contrast the ME-VAE features result in more powerful and informative Spearman correlations with both RPPA pathways and CyCIF (Figure 3.9b). All 10 aggregated features show strong and consistent Spearman correlations, illustrating that the ME-VAE has biological interpretability in both CyCIF and RPPA. Improved correlations illustrate the multi-encoder's applicability for multi-modal integration and comparison by extracting biologically meaningful features.

Biological correlations are validated by looking at representative images for each ligand treatment (Figure 3.9 and Figure 3.10), where the stains shown were selected for their high correlations to the aggregated ME-VAE features or distinct visual patterns. The same patterns observed in the CyCIF correlation table and ME-VAE Z-score expression matrix (Figure 3.9b), are also qualitatively confirmed by visual inspection. For example, S6 expression (ME-VAE feature 0) is high in BMP2+EGF, EGF, and TGFβ+EGF and is low in HGF, OSM, and PBS. Radial CyclinD1 radial slope (ME-VAE aggregated feature 6), as shown in Figure 3.10, is negative in BMP2+EGF, EGF, and TGFβ+EGF, with high stain intensity in the inner compartment and rapid decrease toward the cell perimeter; conversely, HGF, OSM, and PBS show much dimmer CyclinD1 expression in the inner compartment. This pattern is even more clear in the radial HES1 slope (Figure 3.10), where

HGF, OSM, and PBS show a more continuous stain abundance all the way to the cell membrane. Although the RPPA sample size (n=6 independent ligand treated cell populations) is still too small to achieve statistical significance, the correlations between protein markers in CyCIF and RPPA pathways linked by VAE features, are supported by known literature. DAPI expression (ME-VAE aggregated feature 1) is highly correlated to the DNA damage and repair (DDR) pathway, which is expected since DAPI is a maker for DNA expression. A more interesting example (ME-VAE aggregated feature 9) shows a strong correlation between the Stat3 radial slope of distribution and the epithelial-to-mesenchymal transition (EMT) and hormone receptor pathways in RPPA. Prior literature also shows that Stat3 distribution throughout the cell, its translocation to the nucleus, and its cytoplasmic activation are important in the EGF induced epithelial-to-mesenchymal transition pathway[142]. The ME-VAE architecture also extracts patterns when multiple markers play a role; CyclinD1 and p21 (ME-VAE aggregated feature 5) are known in literature to play a joint part in the cell growth pathway[24]. These observations demonstrate a potential application of multi-modal integration using the proposed approach for other single cell image analysis[108].

The ME-VAE can also improve downstream analysis by making the cell populations more easily separable(Figure 3.11) as measured by mean pairwise Tukey p-values and mean effect sizes. For the given MCF10A dataset, the CyCIF markers were chosen with the known ligands and cell populations in mind to highlight differences between the populations and separate them. This results in already decent separability using just CyCIF mean intensity information (Figure 3.11 top).

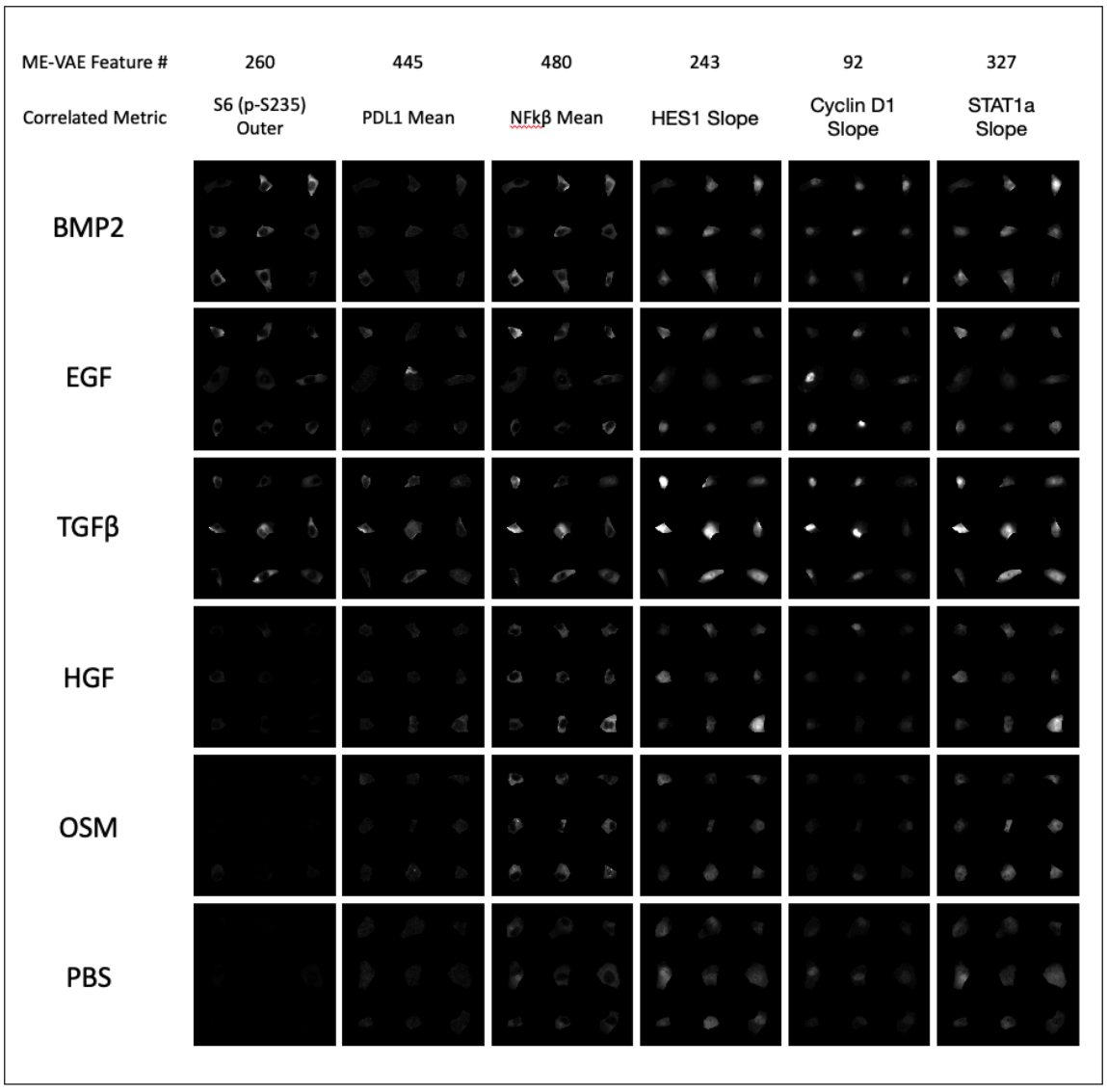


FIGURE 3.10: Representative cell images are shown for each ligand treatment (rows) and are shown using several stains (columns). Each column also includes a VAE number that ties back to the multi-encoder feature that is highly correlated.

That being said, ME-VAE features show lower mean Tukey pairwise p-values indicating a greater average significance in separability, and the effects sizes for those separations are larger (Figure 3.11 bottom). The marker that was an exception to this (shown in the first example) is S6, where the CyCIF mean intensity shows

better separability. Even in this example, however, the multi-encoder's feature is still adequate. It is worth noting that ME-VAE latent space features are encoded in combination to represent even a single stain, so separability can be improved even further when utilizing more than just one feature at a time.

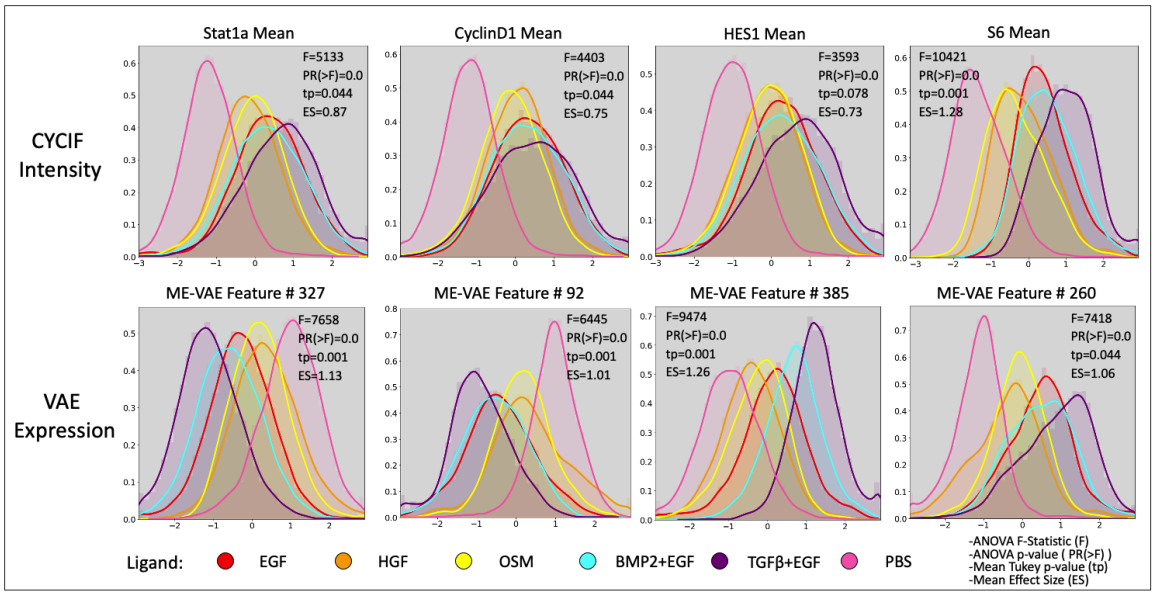


FIGURE 3.11: A two-sided ANOVA was performed for a features and intensities between populations in order to compute the F statistic (F) and p-value (PR(>F)). Subsequently, the mean Tukey-pairwise p-value (tp) across ligands and mean effect size (ES) shown for each feature. ME-VAE features used for comparison were the features with largest correlation to the respective CyCIF marker. This analysis utilized all 73,134 cell images from the MCF10A dataset.

Although aggregated features are useful for integrating data modalities, since they reduce spurious correlations, using the full range of latent features is preferable for clustering populations since aggregation can average out some relevant signal (Figure 3.12). The aggregated features still perform well at separating cell populations and still outperform most stains, however, there is a noticeable reduction in effect size after aggregation.

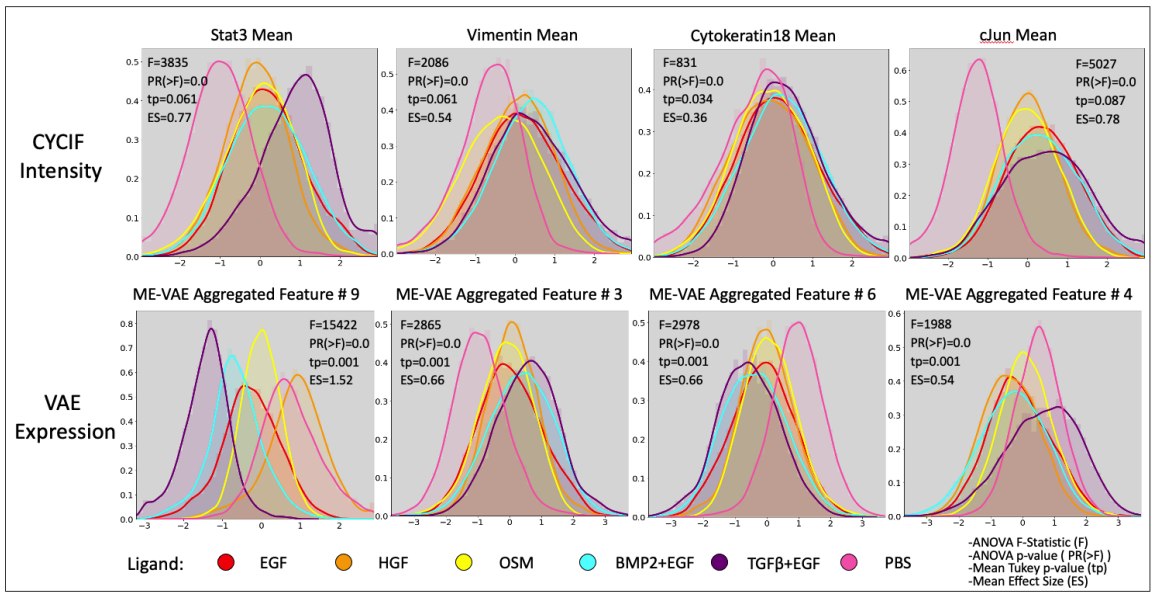


FIGURE 3.12: Density function for several CyCIF and ME-VAE feature pairs. A two-sided ANOVA was performed for a features and intensities between populations in order to compute the F statistic (F) and p-value (PR(>F)). Subsequently, the mean Tukey-pairwise p-value (tp) across ligands and mean effect size (ES) shown for each feature. ME-VAE features used for comparison were the features with largest correlation to the respective CyCIF marker. This analysis utilized all 73,134 cell images from the MCF10A dataset.

3.4 Discussion

Just as it is necessary to pre-process, normalize, and remove unwanted features from single cell RNAseq or RPPA analysis, so too is it necessary to remove uninformative features from single cell imaging analysis in order to extract features of interest. Without this guided feature alignment, VAE applications for single cell image analysis will only reconstruct dominant features while ignoring subtle more informative features (Figure 3.1a/b). By making uninformative features invariable across a dataset using pairs of transformed images in parallel encoding blocks (Figure 3.1g), VAE priority can be shifted to mutually shared, biologically relevant information (Figure 3.3f, Figure 3.6). This results in a more complex and

meaningful latent space.

Feature extraction is important for all downstream analysis and interpretation, but often times naïve metrics are not sufficient to capture biological differences and separate cell populations, especially in datasets where labeled populations are not known beforehand. By separating populations with the ME-VAE, distinct populations and biologically meaningful metrics can be established allowing identification of emergent image properties such as localization and staining pattern (Figure 3.3f, Figure 3.6, Figure 3.7), with increased separability compared to using intensity or morphology information alone (Figure 3.3c/f/g, Figure 3.11). Although a theoretically infinite number of handcrafted naïve features could be crafted to capture more information, the advantage of deep-learning is that it can extract the most important features of an image with limited prior knowledge required. More complex single cell analysis methods such as multimodal integration (Figure 3.9) require a wide range of biologically relevant features. The ME-VAE architecture provides an important step for biological research by linking imaging data to molecular readouts. By employing this architecture to extract a larger range of features and metrics from single cell images, potential applications, such as multi-modal integration using imaging features, become available which were previously restricted due to inadequate cell representations.

To further demonstrate the generalizability of the ME-VAE architecture with a large complex dataset, I applied ME-VAE to a dataset another multiplexed imaging modality, CODEX as described in subsection 3.5.1. The same overall increased performance is observed in the additional dataset of single cell images extracted from CODEX tissue microarrays (TMA)[112] (Figure 3.13), where the Standard

VAE mixes populations and organizes cells primarily based on size. By contrast, the ME-VAE forms distinct clusters with unique expression profiles and is even able to extract cell types with known size differences, for instance, macrophages (as determined by high CD68 expression). In the CODEX dataset, the ME-VAE was only correcting for rotation and polar orientation, since size and shape were considered to be more biologically relevant variables of interest in this setting. This illustrates the ME-VAEs ability to generalize to new modalities, cell types, as well as to tissue data.

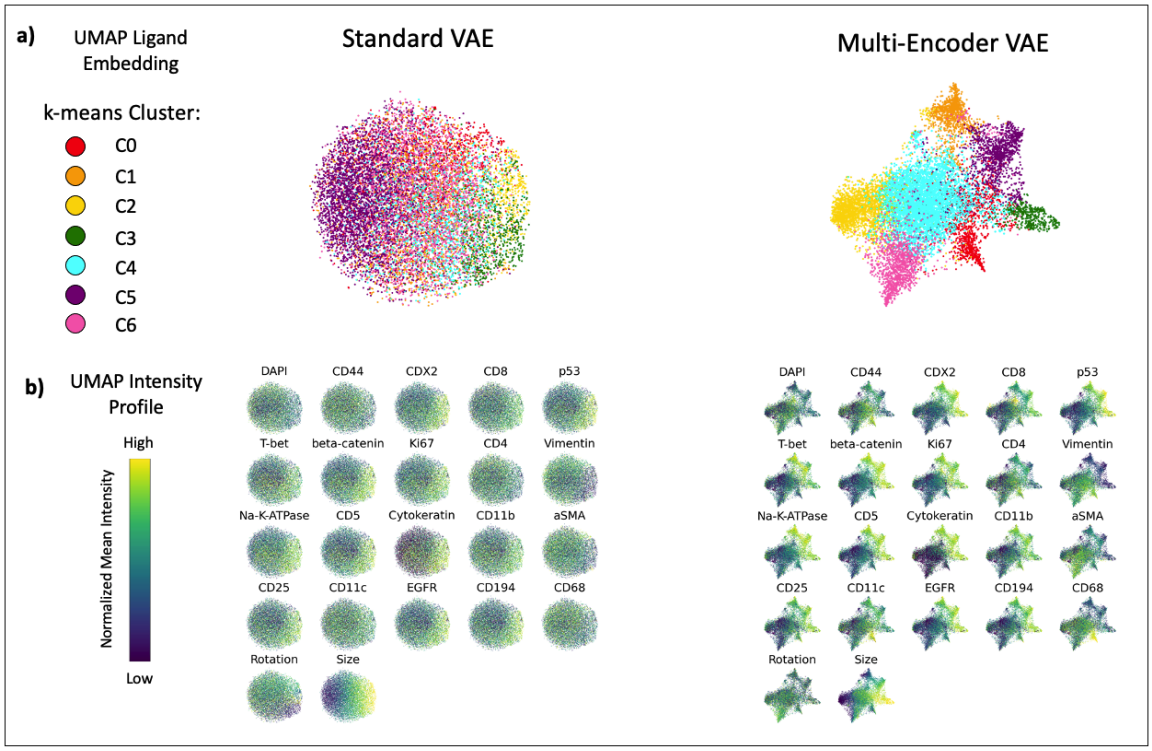


FIGURE 3.13: a) UMAP embeddings for respective VAE encodings, allowing for qualitative visual evaluation of ligand separability. b) Distribution of stain features across UMAP space, colored by intensity. Both models analyzed a dataset of size n=12,229 cells individual cell images.

The simplicity of the multi-encoder design makes it easily incorporated into more

complex deep-learning architectures, such as being augmented with a discriminator to improve reconstruction quality. This methodology is limited by two criteria which the uninformative features must meet: 1) being known so that they can be addressed with a new encoding block and transformed image pair; 2) being a known or inducible transform operation such as rotation, affine, or scale such that a respective randomly transformed image can be generated using the operation. Despite these limitations, the majority of dominant uninformative imaging features are based on known transformations, making the ME-VAE architecture widely applicable. Although researchers have the opportunity with this architecture to add novel non-standard transform features to remove, they will need to verify that the feature being removed is not of biological interest, and the cycle of feature removal might be iterative as new features of disinterest are extracted which weren't previously.

Computationally the model is not significantly larger than a standard VAE or other comparable architectures(Figure 3.1c-g), as the largest amount of additional time is allocated to creating the necessary transformed images, the time for which will vary based on transformation complexity. The increase in computation for the actual architecture is small because it only adds a single encoding block for each undesired feature. Future applications of this architecture will allow complex features such as texture, pattern, and distribution to be extracted from single cell images without the hassle of disentangling dominant uninteresting transform features. Images contain morpho-spatial features not shared by their other single cell counterparts (scRNAseq and RPPA), and by implementing this architecture, the scientific community will be able to analyze these unique image features with the

same robustness as algorithms made for other well-established single cell modalities, with quantitative feature metrics free from the bias of handcrafted feature sets.

3.5 Methods

3.5.1 Datasets

MCF10A cell populations were treated with seven ligands, PBS (control), HGF, OSM, EGF, BMP2+EGF, TGFβ+EGF, and EGF+IFNγ (data from the LINCS Consortium – https://lincs.hms.harvard.edu/mcf10a/)[1]. For this paper I analyzed all but the IFNγ population because initial analysis showed that it was so distinct from other cell populations that even a single marker intensity resulted in decent separability. After 48 hours, cells were fixed and subjected to cyclic immunofluorescence with 23 markers (Table 3.1). The dataset comprises three plates of replicates. On each plate there are three replicates of each ligand in different wells, and in each well 9 different fields of view were taken. Cells were segmented using CellPose segmentation tool[120] using the EGFR and DAPI channels. Stains were normalized using histogram stretching to the 1st and 99th percentiles across intensities for individual cells and across the whole dataset. Image transformations were applied for rotation, polar orientation, and size/shape (Figure 3.14). Rotation is corrected by obtaining the major axis from the binary cell mask, then rotating the image using the Python package OpenCV[13]. Polar orientation was corrected by calculating the angle toward the image's center of mass, then applying a flip/rotation to align the angle using the Python Numpy package[90]. Size/shape was

corrected simultaneously by registering the cell mask to a circle target image (code available here: https://github.com/GelatinFrogs/Cells2Circles). In total, 73,134 cells were processed through this pipeline. When isolating the PBS and TGFβ+EGF populations for the two ligand separation analysis in Results A and B, the sample size was 15,898. All 73,134 cells were analyzed in the full MCF10A analysis, modality integration, and separability test in Results C and D.

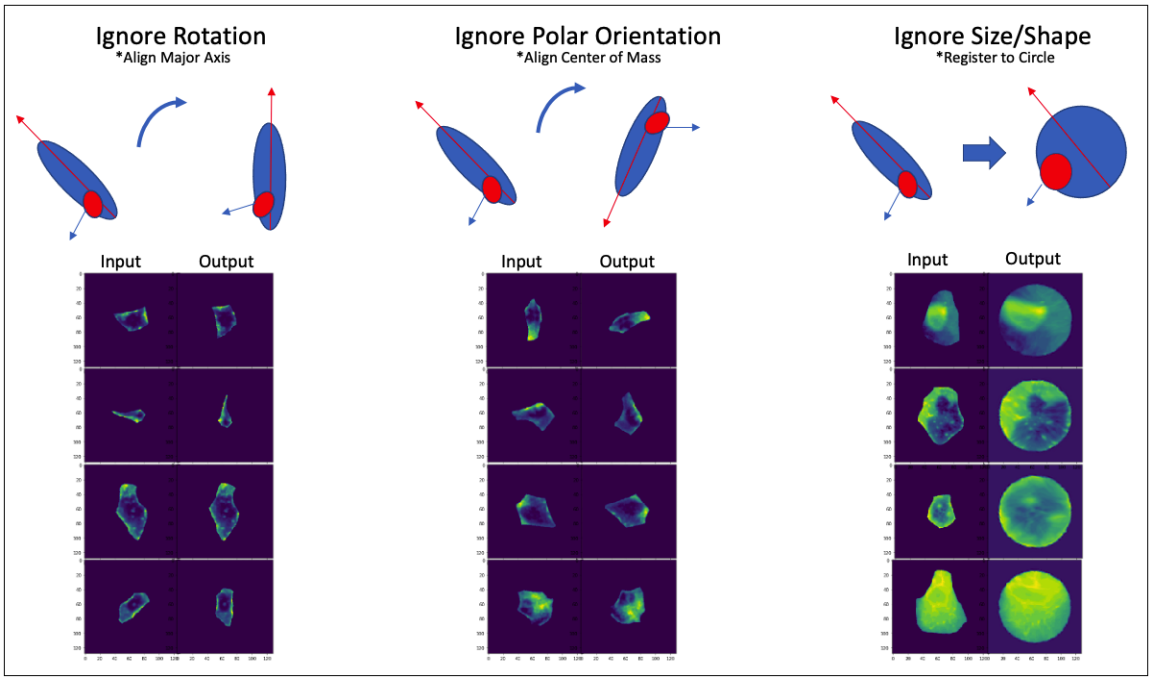


FIGURE 3.14: Examples of image corrections for rotation, polar orientation, and size/shape, shown using EGFR channel of randomly selected images.

A publicly available CODEX dataset[112], was used as a secondary multiplex imaging technology to demonstrate the generalizability of the ME-VAE to other tools, cell types, and to tissue data. The portion of the dataset tested consisted of 8 TMAs from skin and breast cancer. From the full panel of 91 markers, I chose 20 stains that were the least sparse, highest quality, and important for labeling the

full tissue (Table 3.2). I then segmented 12,229 cells from the TMA images using the Hoechst and CD71 channels in Mesmer[40], and normalized using histogram stretching to the 1st and 99th percentiles across the whole dataset.

Channel	Marker
1	HOECHST
2	CD44
3	CDX2
4	CD8
5	p53
6	T-bet
7	beta-catenin
8	Ki67
9	CD4
10	Vimentin
11	Na-K-ATPase
12	CD5
13	Cytokeratin
14	CD11b
15	aSMA
16	CD25
17	CD11c
18	EGFR
19	CD194
20	CD68

TABLE 3.2: CODEX Marker Panel

Bulk Reverse Phase Protein Array (RPPA) was performed by the LINCS consortium[1] in parallel to the CyCIF imaging, on cell populations treated with the same ligands after 48 hours of exposure. The protein array incorporated 295 protein markers. As described by Akbani, R., et al.[3], RPPA data were median-centered and normalized by standard deviation across all samples for each component to obtain the relative protein level. The pathway score is then the sum of the relative protein level of all positive regulatory components minus that of negative regulatory components in a particular pathway. Pathway members and weights were developed through literature review. Pathways were used instead of individual proteins because the large number of proteins would decrease the significance of correlations. Despite the available bulk RPPA dataset saving a smaller sample size

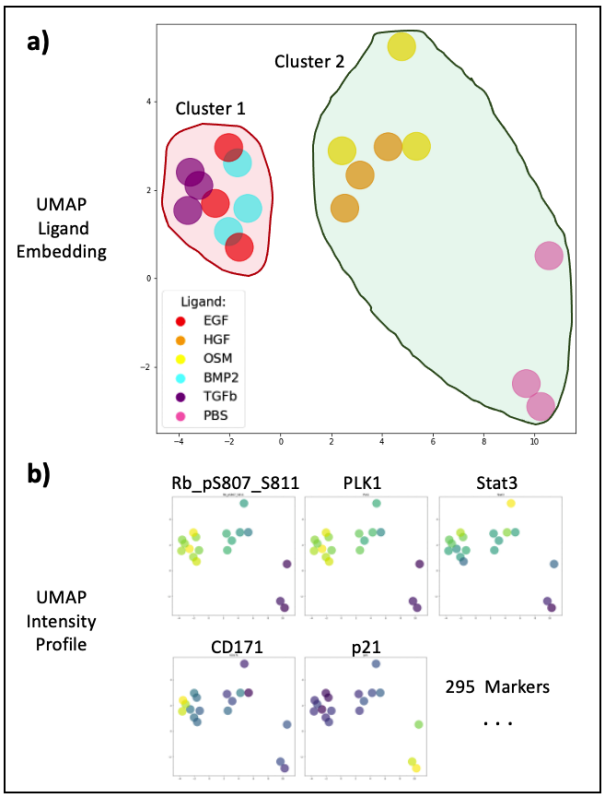


FIGURE 3.15: a) Independent analysis of the Bulk RPPA dataset shows distinct clustering of ligand populations in UMAP embeddings space where b) selected markers show clear patterns of distribution between the clusters.

than the single cell CyCIF dataset, it was chosen as the secondary modality because similar ligand separation and cluster patterns were observed in both modalities, indicating an overlap in the information each contains (Figure 3.15).

For correlation to CyCIF and RPPA pathways, the VAE latent space was restricted to smaller sets of aggregated features. These aggregated features were made using self-correlation of VAE features across individual cell metrics and averaging the VAE features for resulting hierarchical clusters (Figure 3.9 and Figure 3.8). This was done to reduce the feature dimensionality and reduce spurious correlations in the biological findings. Representative images for each cluster were done by

finding cells with a high average expression for all features within the cluster. For other analyses of VAE features comparing VAE separability to CyCIF expression and interpreting image feature space, ME-VAE encoding features were restricted to 18 single features for each. The dimension of 18 was chosen because it is the number of mutual markers between the RPPA and CyCIF datasets. Explanatory features were chosen from the VAE encodings such that the inter-cluster variability was maximized and the intra-cluster variability was minimized using the following equation:

$$FeatureScore = Var_{all} - \frac{\sum_{i=cluster} Var_{c_i}}{\#ofClusters} \tag{3.1}$$

3.5.2 VAE models

To allow for fair comparison, the structure of the encoder and decoder blocks were kept consistent between networks, and the same latent dimension was used for all models for a given dataset (64 for the 1-channel dataset, 512 for the 23-channel dataset). Both encoder and decoder blocks consist of three layers, and all layers utilize a relu activation except the final output, which uses sigmoid activation. All models were trained for 10 epochs (determined by identifying the loss function plateau) on the NVIDIA P100 with 100GB of RAM and 100GB of disc space, but the ME-VAE architecture can work on any NVIDIA GPU.

A standard VAE with matching pairs of single cell images was used to establish baseline performance (Figure 3.1c). Standard VAEs utilized the standard Evidence Lower Bound (ELBO) loss format characterized by reconstruction and Kullback–Leibler (KL) divergence terms. I used a Binary Cross Entropy loss (BCE) as the reconstruction term for all VAEs tested here to keep the comparisons fair and consistent. Put together, the standard VAE loss used was:

$$L_{StandardVAE} = BCE(x, p(z)) - KL[q(z|x)||p(z)] \tag{3.2}$$

where q represents the encoder and p represents the decoder as described in Kingma et al.'s initial VAE paper[58]. Here x represents the unadjusted input image and z represents the latent space.

By using an image randomly transformed with respect to a dominant feature as the input and controlling for the same uninformative feature in the output image (Figure 3.1d), the model can self-supervise the transformation and will only encode novel features since the controlled features (such as rotation) no longer aid reconstruction:

$$L_{OutputCorrectedVAE} = BCE(x', p(T^{-1}(x'))) - KL[q(z|x')||p(z)] \tag{3.3}$$

where x′ represents an image that has been transformed with a known transformation to remove one or more uninformative features and $T^{-1}(*)$ represents a transformation of the controlled image to create a dominant uninformative feature at a random degree.

The proposed multi-encoder architecture uses multiple transformed inputs with separate encoder blocks, where each block controls for a separate uninformative feature, and a single decoder block uses the shared latent space (combined by multiplication to emphasize mutual information) for reconstruction (Figure 3.1f). To accommodate the multiple encoders in the loss, the KL term is replaced with a summation of all KL divergences for each individual latent space, which is then divided by the total number of encoders (n):

$$L_{ME-VAE} = BCE(x', p(z_{all})) - \frac{1}{n}(\sum_{1}^{n} KL[q_i(z_i|T_i^{-1}(x'))||p(z_{all})]) \tag{3.4}$$

where each encoder's (q_i) individual latent space (z_i) is combined in an elementwise multiplication layer to create a mutual latent space (z_{all}) and $T_i^{-1}(*)$ represents a different random transformation for individual uninformative features such as rotation, polar orientation, size, shape, etc. respectively. The shared latent space of the multi-encoder forces the deep learning model to encode features that are shared between each transformation, reinforcing the shared mutual information and eliminating the non-shared transformational information. A base implementation of the ME-VAE architecture can be found here: https://github.com/GelatinFrogs/ME-VAE_Architecture. The multi-encoder architecture allows for image pairs to be randomly transformed, which can act as a balancing agent for imbalanced features. Furthermore, the corrected outputs serve as a weakly self-supervised signal for the model. With the extra information from the additional inputs, the model is able to overcome more complex transformations that failed in the corrected output architecture in

Figure 3.1d, when multiple corrections are attempted. Paired images also serve one additional benefit of allowing for features to be retained in parallel encoders that might be lost due to artifacts in other corrections, i.e. artifacts within a polarity correction encoder will not be present in a rotation correction encoder.

The β-VAE[44] makes a small but significant change to the standard ELBO loss function of the Standard VAE by adding an adjustable hyperparameter to the loss function:

$$L_{\beta-VAE} = BCE(x, p(z)) - \beta \cdot KL(q(z|x)||q(z)) \tag{3.5}$$

where the β applies varying amounts of priority to the KL regularization term. As described in the β-VAE paper[44], this forces the VAE to separate features into more interpretable format where each component corresponds to a specific feature. One downside of this method is that by shifting priority to the regularization term is that it causes the model to produce poorer quality reconstructions since less priority is placed on the reconstruction term. Another downside is that the β hyperparameter can be difficult to tune properly, and a different β value will be optimal for different datasets, image sizes, and latent dimensions (Figure 3.4).

A final architecture tested was the invariant conditional autoencoder (C-VAE), which injects the quantified class/values of interest into the decoder. The improvement this architecture makes upon the ELBO loss used by standard VAEs and conditional VAEs is the addition of a conditional and marginal KL regularization term that operates similar to a Maximum Mean Discrepancy penalty in that it

"encourages the statistical moments of each [latent space] to be the same over the varying values of c"[86]:

$$L_{C-VAE} = -KL[q(z|x)||p(z)] - \lambda \cdot KL[q(z|x)||q(z)] + (1+\lambda)BCE(x, p(z)) \quad (3.6)$$

where λ is a hyperparameter which during this experiment was 1. The invariant conditional VAE was adapted based on the paper by Moyer et al.[86] and code available from the author's Github and tutorials (https://github.com/dcmoyer/invariance-tutorial/blob/master/tutorial.ipynb). In our implementation, I used the quantified values of rotation angle, polar orientation, and size as the C inputs such that the latent space would hopefully be invariant to those features. Similar to our proposed approach, the uninformative features of interest must be known in this method since the C values are input into the model.

Additionally, in the reduced dataset the architectures are compared to classically extracted intensity and morphology features using scikit-image's regionprops package[138]. The classical feature dataset is defined by 58 properties (Table 3.3) extracted from each single channel cell images. I included all properties I could for single channel images, but left out orientation in order to show that rotation angle is still captured through other properties even when it is not explicitly an extracted feature.

#	Property		#	Property
1	area		31	inertia_tensor-0-0
2	moments_central-0-0		32	inertia_tensor-0-1
3	moments_central-0-1		33	inertia_tensor-1-0
4	moments_central-0-2		34	inertia_tensor-1-1
5	moments_central-0-3		35	inertia_tensor_eigvals-0
6	moments_central-1-0		36	inertia_tensor_eigvals-1
7	moments_central-1-1		37	major_axis_length
8	moments_central-1-2		38	max_intensity
9	moments_central-1-3		39	mean_intensity
10	moments_central-2-0		40	minor_axis_length
11	moments_central-2-1		41	moments-0-0
12	moments_central-2-2		42	moments-0-1
13	moments_central-2-3		43	moments-0-2
14	moments_central-3-0		44	moments-0-3
15	moments_central-3-1		45	moments-1-0
16	moments_central-3-2		46	moments-1-1
17	moments_central-3-3		47	moments-1-2
18	centroid-0		48	moments-1-3
19	centroid-1		49	moments-2-0
20	ecccentricity		50	moments-2-1
21	euler_number		51	moments-2-2
22	extent		52	moments-2-3
23	ferret_diameter_max		53	moments-3-0
24	moments_hu-0		54	moments-3-1
25	moments_hu-1		55	moments-3-2
26	moments_hu-2		56	moments-3-3
27	moments_hu-3		57	perimeter
28	moments_hu-4		58	solidity
29	moments_hu-5		–	–
30	moments_hu-6		–	–

TABLE 3.3: RegionProps classical features list

3.5.3 Evaluations metrics

In order to evaluate the model's ability to separate labeled cell populations, k-means clustering was applied to the encoding spaces using sklearn[96]. Cluster purity was then calculated by taking the percentage of the largest population for each cluster. UMAP embeddings were calculated using the UMAP Python package[85]. Biological metrics were calculated to give VAE encoding features biological grounding (Figure 3.5). Circularized cells were used for calculation because it made compartmentalization of the cell more consistent and uniform. Mean intensities were calculated for inner, middle, outer, and whole cell compartments. To

calculate the radial slope, the mean intensity was taken from each radius of the circularized cell, then the linear regression of the series was calculated using the scipy.stats package[136] in Python. The slope of the calculated linear regression was used as the metric and the intercept was ignored. Self-correlations between VAE features were performed using Spearman correlation and clustering was done in seaborn clustermap.[144] Clustermaps using hierarchical clusters were calculated using the function's default method (Euclidean). Representative cluster images were chosen based on high expression of the cluster's respective VAE features. RPPA pathways activity scores, VAE features, and biological metrics were all standardized prior to analyses using the sklearn StandardScaler function[96] in Python. Correlations between RPPA pathway activities and VAE encodings and between CyCIF and VAE encodings were both calculated using the Spearman correlation. To test for separability (Figure 3.11 and Figure 3.12), features were first tested using type 2 ANOVA with the Python implementation of anova_lm from statsmodels[113] for the default F-statistic, all of which proved significant. Subsequently, the post-hoc pairwise Tukey p-test was used to calculate the significance and effect size for each ligand pair. The mean p-value and effect size were reported to illustrate average separability.

3.6 Code availability

For reproducibility, I share the code with precise implementation, further details describing variables and equations, as well as shared trained models with parameters in Github. All ME-VAE code is available on Github

(https://github.com/GelatinFrogs/ME-VAE_Architecture)

3.7 Data availability

CyCIF and RPPA Data is publicly available through the LINCS Consortium: (https://lincs.hms.harvard.edu/mcf10a/) CODEX Data is available online (https://doi.org/10.7937/tcia.2020.fqn0-0326)

Chapter 4

Guiding multiplex imaging with stain propagation, region selection, and panel reduction

You are still guided by your own expectations...
While you are still plotting, do you think you can really be guided in what to do?

Readings from Chuang Tzu

4.1 Abstract

Multiplex tissue imaging platforms (MTIs) generate large amounts of data with unprecedented scale, resolution, and depth. Although this bulk of information presents the opportunity for many novel discoveries, it also comes with many new challenges concerning how to tackle the creation and interpretation of the

data most efficiently. Tissue based sampling and diagnosis is defined as extraction of information from certain limited spaces and the diagnostic significance of a certain object. Many MTIs make the assumption that tissue microarrays (TMAs) containing small core samples of 2-dimensional (2D) tissue sections are a good approximation of bulk tumor even though tumors are not 2D. However, emerging whole slide imaging (WSI) or 3D tumor atlases which employ MTIs like cyclic immunofluorescence (CyCIF) strongly challenge this assumption. In spite of the additional insight gathered by measuring the tumor microenvironment in WSI or 3D, it can be prohibitively expensive and time consuming to process tens or hundreds of tissue sections with CyCIF. Even when resources are not limited, the criteria for region-of-interest (ROI) selection in tissues for downstream analysis remain largely qualitative and subjective as stratified sampling requires the knowledge of objects and evaluation of their features. Despite the fact TMAs fail to adequately approximate whole tissue features, a theoretical subsampling of tissue exists that can best represent the tumor in the whole slide image. Similarly, MTIs like CyCIF utilize large panels of markers that attempt to gather as much information as possible, but increasing the number of stains does come with the downsides of increased autofluorescence and tissue degradation. Just as with spatial sampling, there also exists a theoretical subsampling of markers that is able to recreate the same information as a full panel; therefore, removing the self-correlating information with such a subset would increase the efficiency of the imaging process and maximize the information collected. To address these challenges, I propose two deep learning approaches to learn multi-modal image translation: 1) generative modeling

approach to reconstruct 3D CyCIF representation and 2) co-embedding CyCIF image and Hematoxylin and Eosin (H&E) section to learn multi-mappings by a cross-domain translation for minimum representative ROI selection. I demonstrate that generative modeling enables a 3D virtual CyCIF reconstruction of a colorectal cancer specimen given a small subset of the imaging data at training time. By co-embedding histology and MTI features, I propose a simple convex optimization for objective ROI selection. I demonstrate potential application of ROI selection and the efficiency of its performance with respect to cellular heterogeneity. Finally, I test and utilize several embedding and reconstruction strategies to determine the best method for selecting an optimized panel set.

4.2 Introduction

Cancers are complex diseases that operate at multiple biological scales—from atom to organism—and the purview of cancer systems biology is to integrate information between scales to derive insight into their mechanisms and therapeutic vulnerabilities. From this holistic perspective, the field has come to appreciate that the spatial context of the tumor microenvironment in intact tissues not only enables a more granular definition of disease, but also the design of more personalized and effective therapies[75]. This has been spurred by an increased understanding that solid tumors are complex ecosystems including stromal barriers imposed by tissue architecture[54] and infiltrating immune cells in the surrounding stroma[99]. This has motivated the National Cancer Institute's Human Tumor Atlas Network

(HTAN) to begin charting 3D tissue atlases which capture the multiscale organizations and interactions of immune, tumor, and stromal cells in their anatomically native states[104]. The HTAN-SARDANA[67] is one such atlas which aimed to deeply characterize the architecture of a single colorectal cancer (CRC) specimen via histology and a spatial context-preserving multiplexed imaging platform called cyclic immunofluorescence (CyCIF)[65].

Histology is an essential component of the clinical management of cancer. For around 150 years, pathologists have interrogated thin sections of tissue stained with hematoxylin and eosin (H&E) to determine the morphological correlates of cancer grade, stage, and prognosis. However, this essentially 2D representation of tissue is a relatively poor representation of tissues like prostate, pancreas, breast, and colon which have highly convoluted 3D ductal structures [71, 56, 18, 66]. Since 2D whole slide imaging of a 3D specimen might not be representative, 2D analyses using biased downsampling or the small fields of view afforded by tissue microarrays (TMAs) suffer further due to subsampling issues[67]. Moreover, histology alone lacks the molecular specificity to unequivocally determine the identity and function of cells in tissue. In contrast, CyCIF enables the co-labelling of tens of markers in tissue and can broadly characterize the tumor, immune, and stromal compartments. By coupling histology and CyCIF in the same specimen, the HTAN-SARDANA atlas integrates both top-down (pathology-driven) and bottom-up (single cell phenotype-driven) perspectives of CRC and provides a framework for the charting of 3D atlases for other cancers[67].

In spite of these advances, 3D multiplexed imaging atlases and 2D whole slide

multiplexed imaging with large cohorts both require a tremendous amount of resources and effort to build. For the HTAN-SARDANA atlas, a single CRC specimen was serially sectioned and processed yielding 22 H&E slides interleaved with 25 CyCIF slides, with the CyCIF slides taking days to process due to the cycles of antibody incubation. To build the breast cancer atlas in [18], a single specimen was serially sectioned and processed into 156 slides which were characterized using imaging mass cytometry, which enables simultaneous labeling of 40 antigens with a single incubation step, but has relatively limited spatial scope (50 µm x 50 µm x 50 µm) compared to CyCIF. To build the pancreas cancer atlas in [56], specimens were serially sectioned and processed into over 1,000 H&E slides, some of which had histological regions of interest labeled through a laborious and subjective manual annotation process. These annotations were used as training data for a deep learning segmentation model which was used to fully reconstruct the labeled classes of the 3D specimen at the pixel level with high accuracy, but this approach is restricted by the limited and predefined annotation classes.

To address this challenge, I extend a virtual staining paradigm into the third dimension by deploying it on the coupled H&E and CyCIF image data from the HTAN-SARDANA atlas of CRC. There have been several applications at stain prediction within the limited context of a single two dimensional tissue section[16, 23, 91, 61]. I have also previously demonstrated methods for predicting virtual IF stains based on H&E-stained tissue (SHIFT: Speedy Histological-to-ImmunoFluorescent Translation)[133], wherein I use spatially-registered H&E and immunofluorescence (IF) data and generative deep learning to model the correspondences between these imaging modes and compute near-real time

virtual IF stains conditioned on H&E-stained tissue alone. SHIFT is a deep learning architecture made from generator and descriminator models that works to create stylistically realistic stain images from H&E. From a biological perspective, these data and approaches allow us to ask which markers in an IF panel have a quantifiable histological signature, what that signature might be, and a means to estimate the distribution of markers in histological images for which such a signature exists. From an application perspective, the approach could be useful for automated compartment labeling in 3D tissues labeled with highly-standardized and low-cost histological stains. I demonstrate that what generative models learn from less than 5% of coupled H&E and CyCIF images is sufficient to generate a virtual 3D CyCIF reconstruction of the whole CRC specimen and that quantitative endpoints derived from real and virtual CyCIF images are highly correlated.

In order to reduce the burden and complexity of multiplex imaging on whole slide images (WSIs), TMAs are often used to sample small sections of the tissue for analysis. Although these TMAs have become a staple of analytics over the past decade, they come with many drawbacks and are prone to substantial bias, often introducing sampling errors and shifts in the expected content which fail to accurately capture the true heterogeneity and spatial distributions found in WSIs[88]. In order to overcome this sampling bias, a significantly large number of TMA cores would need to be taken[63], but increasing the size of the randomly sampled TMA cores also shows little to no effect on improving their representativeness[89]. It is necessary to intelligently sample regions for TMAs, but without a method to quantify biological content beforehand, intelligent sampling is estimated from histological

appearance alone. If regions of WSIs could be quantitatively described prior to analysis, TMA cores could subsequently be taken based on which regions of the image were most representative of the whole slide.

Methods for selecting important regions have been performed in H&E previously[95], but most attempt only to capture high level tissue features and do not attempt to capture cell type information that would be useful for inferring CyCIF information. The ability to capture such expression based information would be necessary to select regions that are going to be important for subsequent staining and analysis. As a method for select core-like regions similar to doing a TMA virtually, I utilize a shared representation between H&E and CyCIF to quantitatively identify representative samples that will serve as the optimal regions of interest (ROIs). Using the principles of SHIFT[133], here I propose a cross-domain autoencoder (XAE) image translation architecture which after training can assign regional descriptors to image tiles that contain the cell type information of CyCIF based solely on the H&E image. By formulating a simple convex optimization problem, these tile-based descriptors can be used to select small regions that are representative of the whole slide image with a minimum number of ROIs. I demonstrate that the XAE architecture is able to adequately represent biological information and that the minimum set of ROIs is more representative of whole slide biology than random sampling or biased manual ROI selection.

Even within these ROI sections, panels must be chosen intelligently and with purpose because each additional round of staining comes with additional deleterious tissue effects that negatively effect the resulting images and downstream analysis.

Some examples of this include increased levels of autofluorescence, caused by the natural emission of light by the biological structures and proteins in the normal blocking serum[53], as well as the degradation of tissue that gets washed away between staining rounds. Current algorithmic methods to select panels do not operate at the large scale of multiplex imaging data, and instead focus on selecting panels that optimize the physical restrictions of imaging, such as the overlap of wavelengths between markers[135]. When selecting a panel, however, experts pay considerable attention to the biology of the disease in order to capture specific features of interest, but it is not always possible for experts to know the full extent of marker co-expression, co-localization, and predictability. Just as it is possible to use deep learning architectures to predict CyCIF information from features in H&E alone, so too is it possible to predict shared information from one marker to another. By selecting an idealized subsample of markers, a deep learning model can be trained to predict the same information as a full dataset with fewer rounds of staining. Here I evaluate several methods of subsample marker selection and demonstrate their ability to reconstruct the full panel's information.

Data generation and analysis are expensive and time consuming, and choices often have to be made to reduce the scope of experiments, which will result in the loss of potential vital information. Here I explore three methods to maximize the amount of information retained using computational methods at different points of the multiplex imaging pipeline as shown in Figure 4.1): 1) 3D stain propagation using SHIFT for stain prediction 2) guided ROI selection using a cross-domain autoencoder 3) Optimized panel selection to maximize the imputation of absent markers.

Using these computational methods, it will be possible to obtain the most information possible from multiplex imaging while reducing the amount of staining, tissue, markers.

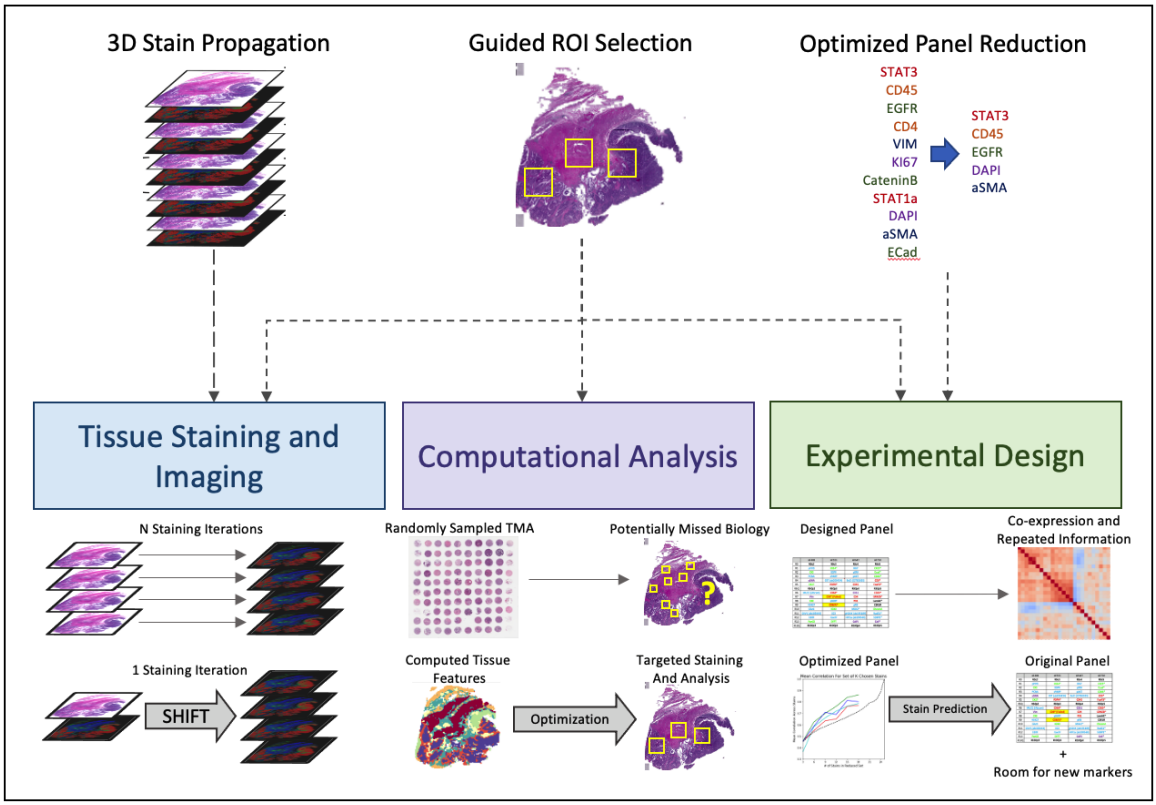


FIGURE 4.1: All three methods discussed work to improve the multiplex imaging pipeline from the basics of experimental design, to imaging, to downstream analysis by maximizing the amount of information obtained with computational inference and guidance using minimal data as input. 3D stain propagation reduces the number sections that need to be stained and imaged by learning to infer stain distributions from H&E. Guided ROI selection allows for a smaller amount of tissue to be stained and analyzed while maximizing the relevant biology by targeting computed biological features. Panel reduction allows for researchers to infer the information of a full panel with computationally selected markers, reducing staining rounds and opening space for new markers on the panel.

4.3 Results

4.3.1 3D stain propagation using SHIFT

A. Preprocessing steps for spatially registered H&E and IF images

Spatially registered H&E and IF images are a requirement for SHIFT model[133] training and evaluation. To register the H&E and CyCIF data for this task, I begin with sequential registration of the H&E stack beginning from the middle sections and propagating to outer sections (section 4.5, Figure 4.2A/B). I then co-register ROIs of adjacent H&E and CyCIF images using their respective nuclear masks for a finer local registration of the adjacent sections.

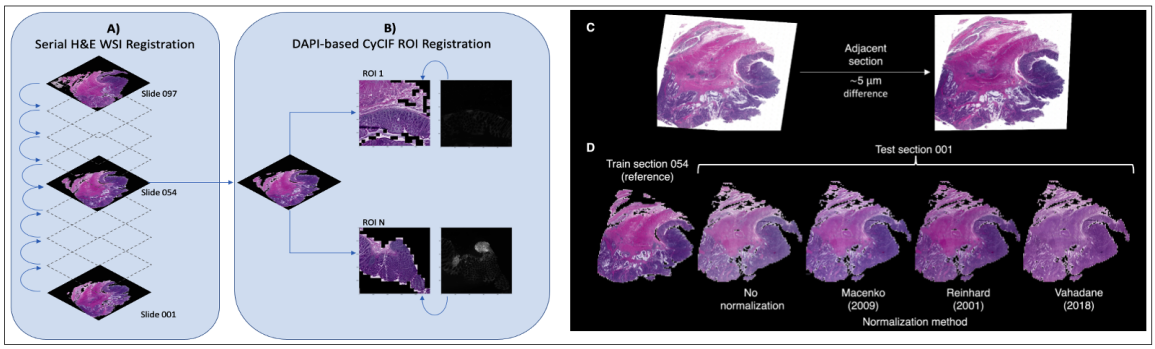


FIGURE 4.2: A) To register H&E in the three-dimensional setting, I sequentially registered all slides to the center using the transforms propagated from previous layers. B) CyCIF was then finely registered to the adjacent H&E images at the ROI level to maximize single cell level correspondence. Registration of CyCIF and H&E was performed using binarized DAPI and thresholded H&E to align nuclei. C) Tissue sections are subject to technical variability in stain intensity, even between adjacent sections that are separated by only 5 µm. D) Representative results of H&E stain normalization. The stain intensity distribution of the test section 001 is transformed to match that of the reference section 054 which was used for SHIFT model training.

Before SHIFT model training could begin, I had to account for the section-to-section variability in H&E stain intensity, which helps to ensure a model trained on one H&E section generalizes well to the other sections. Using the training H&E

section (middle section as shown in Figure 4.2A) as reference, I tried several stain normalization methods for outer testing sections[98, 134, 76], and found that the Reinhard method worked best at normalizing stain intensities to the reference by qualitative comparison (Figure 4.2C/D). This result was consistent with a quantitative comparison that found the Reinhard method conferred better generalizability to DL models in an analogous digital pathology application[127].

B. Image-to-image translation for 3D virtual CyCIF reconstruction

With spatially registered H&E and CyCIF data, I generated a virtual 3D CyCIF reconstruction in an effort to measure how faithfully I can characterize the full SARDANA dataset with virtual IF staining by learning from only one pair of adjacent H&E and real CyCIF sections. First, the middle pair of H&E and CyCIF sections was selected for training SHIFT models under the assumption that they are a good representation of the tissue on either side of the sample block. This assumption is supported by the initial HTAN-SARDANA study[67], where the authors conclude that 2D whole slide imaging of a 3D specimen does not, in general, suffer from the subsampling issue associated with TMAs or small fields of view.

I then decompose the WSIs into thousands of pairs of matching H&E and IF image tiles, and use those to train a generative adversarial network (GAN) to synthesize virtual IF tiles conditioned on H&E tiles[133]. Briefly, the generator network of the model is responsible for synthesizing virtual IF images conditioned on H&E images, and the discriminator network is responsible for quality assurance of the virtual IF images synthesized by the generator as shown in Figure 4.3A. Once trained on the middle-most sections, the model can then be tested using the tiles from the

held-out H&E sections. The generated IF images are then compared with the real CyCIF images to evaluate performance. Importantly, a virtual IF image is conditioned on an H&E section, and there is natural variation between it and its adjacent real IF section 5 µm away, which complicates pixel-wise evaluation of model accuracy.

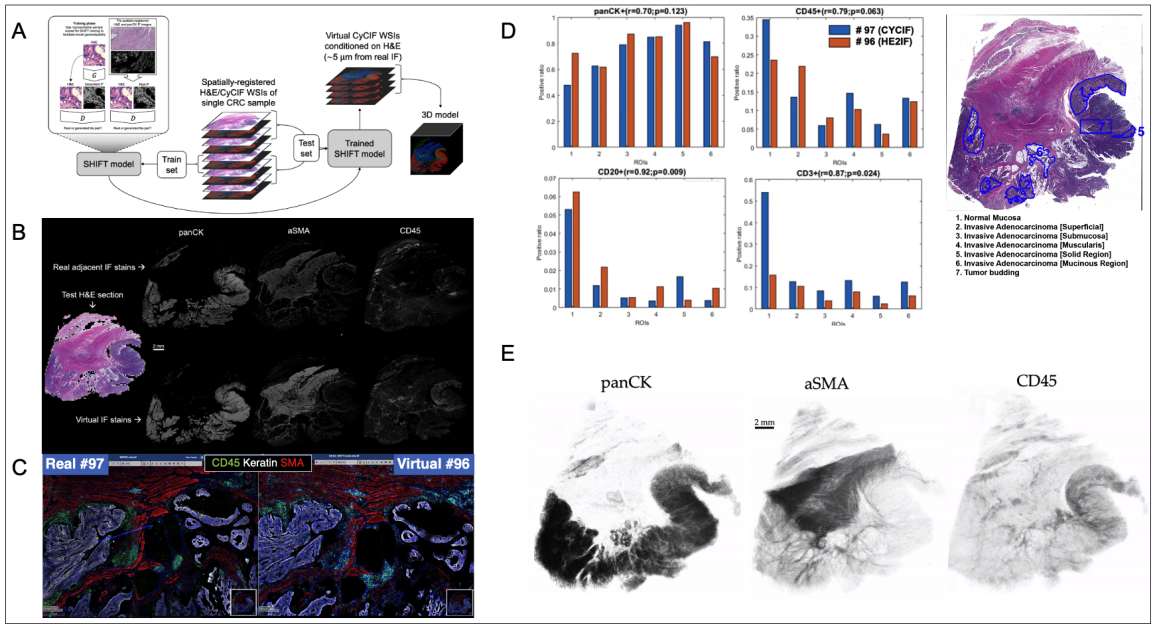


FIGURE 4.3: (A) Extending SHIFT to 3D using adjacent spatially-registered H&E/CyCIF WSIs from a single CRC sample. (B) WSI virtual staining result. Models trained to predict single-channel CyCIF images conditioned on the H&E/CyCIF training sections were applied to H&E test section 096 to generate virtual stain WSIs for the markers panCK, aSMA, and CD45. The input H&E test section is shown at left, and the real and virtual CyCIF WSIs are shown in the rows above and below, respectively, for ease in comparison. (C) qualitative comparison of real and virtual staining for the markers panCK, aSMA and CD45 in the selected region. (D) Quantitative comparison of ROI cell composition correlation between real. For each of the ROIs, the positive ratio of cells for each of panCK, CD45, CD20, and CD3 are calculated using the same workflow and displayed for either real or virtual CyCIF WSIs. Pearson's correlations and p-values describing the association between positive ratios derived from real and virtual CyCIF WSIs for each marker are indicated above each bar plot. (E) 3D virtual stain volumes conditioned on held-out H&E test sections visualized by 3D Slicer[30].

I trained individual SHIFT models to predict single CyCIF channels conditioned

on H&E inputs from the central H&E/CyCIF training sections 053/054 (Figure 4.3A). Representative test results from the application of trained SHIFT models on H&E/CyCIF test sections 096/097 (far from the middle section, i.e., training section) are shown in Figure 4.3B/C. These qualitative results indicated that the SHIFT models were fitting well to the training sections, and the representations learned were useful for extension to held-out test sections.

The virtual CyCIF images generated by SHIFT models are conditioned on H&E sections which are 5 µm adjacent to the real CyCIF sections, so the cellular contents are slightly different between sections and images. Recognizing that this would hamper pixel-wise comparisons between the real and virtual CyCIF images[133, 20], I estimated an upper bound on SHIFT performance by measuring the concordance between nuclear content from the adjacent sections of the H&E/CyCIF test sections 096/097 (Figure 4.4).

The test sections were first subdivided into 135 non-overlapping ROIs and each ROI was locally registered to improve the alignment of H&E and CyCIF image content, then I measured the Dice coefficient of nuclear masks derived from the H&E and DAPI images from each ROI (Figure 4.4A). I used the Dice coefficient for each ROI as a compensation factor when evaluating the quality of the virtual stains for each ROI by dividing raw quality scores by the Dice coefficients corresponding to each ROI. Virtual CyCIF image quality was evaluated using structural similarity index measure (SSIM), which is established as a metric for assessing virtual stain quality[133, 100, 101]. Median SSIM score following compensation by the upper bound ranged from 0.36 for CD20 up to 0.89 for aSMA. This result suggested that there was significant room for improvement for some SHIFT models, but with the

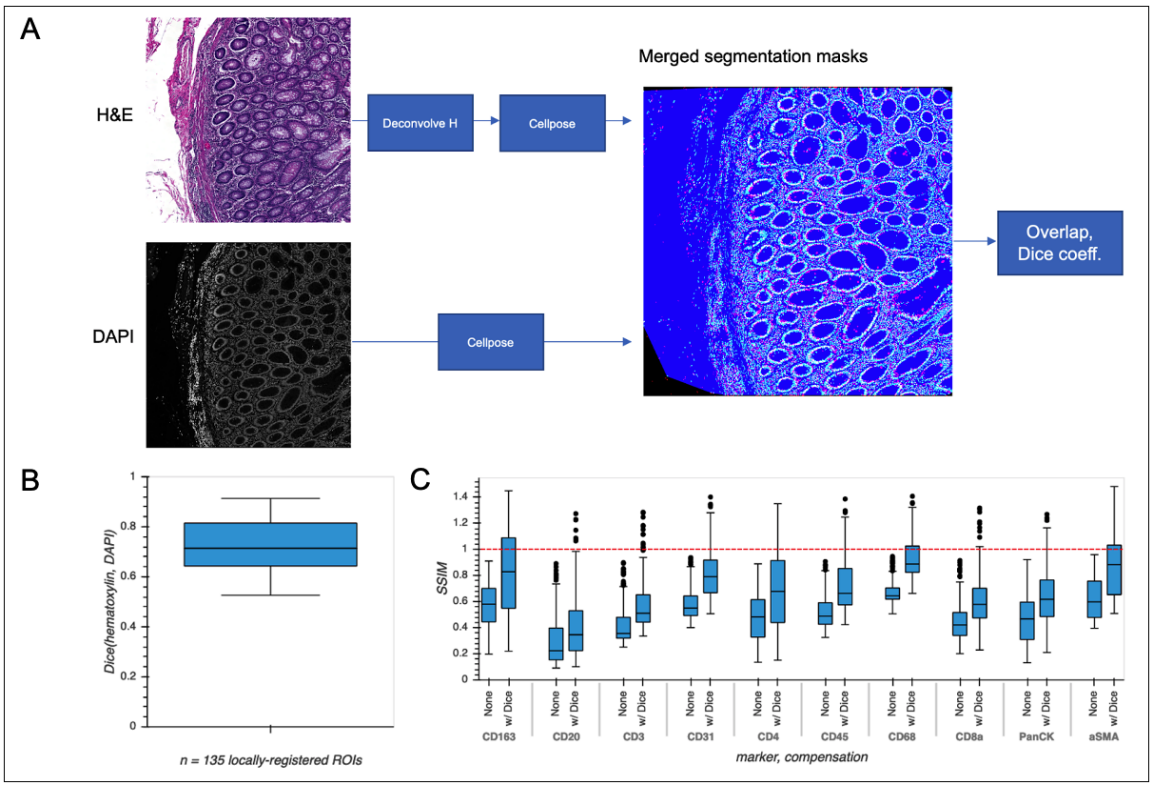


FIGURE 4.4: (A) The Dice coefficients describing the overlap of nuclear masks from ROIs of adjacent sections were used as compensation factors for evaluating virtual stains. (B) Boxplot describing the distribution of Dice coefficients of the 135 locally-registered ROIs from H&E/CyCIF test sections 096/097. (C) Boxplots describing the distributions of structural similarity (SSIM) of real vs. virtual CyCIF ROIs over the 135 locally-registered ROIs from H&E/CyCIF test sections. The red dotted line indicates the unity line describing Dice-compensated SSIM.

adequately performing channels, the virtual images will still prove useful to researchers, since SSIM is sensitive to slight differences in image contrast which may not significantly affect downstream processing and interpretation[133].

To test this, I quantified the positive cell ratio for multiple markers in each of pathologist-annotated 6 ROIs in H&E test section 096 using either real or virtual

CyCIF images (Figure 4.3D), which assesses how such an endpoint might be impacted when using virtual images which may or may not be of high quality with respect to SSIM (Figure 4.4). In spite of the adjacency complication explained above, there was substantial correlation between positive cell ratios using real and virtual CyCIF images, suggesting that virtual images could be used in place of real without significantly affecting some downstream endpoints. Having established the fitness of the SHIFT models, I performed a full virtual 3D reconstruction of the CyCIF images by passing all held-out H&E test sections to the SHIFT models trained on the H&E/CyCIF training sections (Figure 4.3E).

I also assessed the value added by the discriminator network of the GAN by training models without it, leaving the generator network to learn the virtual panCK stain alone (Figure 4.5). I found that while the generator-only virtual panCK stain has good localization, it lacks the naturalistic texture of the real and GAN-generated virtual stains, which highlights the compromise of a more efficient and portable generator-only model.

4.3.2 Guided region-of-interest selection

A. Shared latent representation via embedding of CyCIF images on H&E image

3D Virtual staining is enabled through the rich latent representations that generative models are capable of learning from paired H&E and CyCIF image data. I hypothesized that these latent representations could be useful for the related and unsolved problem of objective ROI selection. If ROI selection for targeted CyCIF

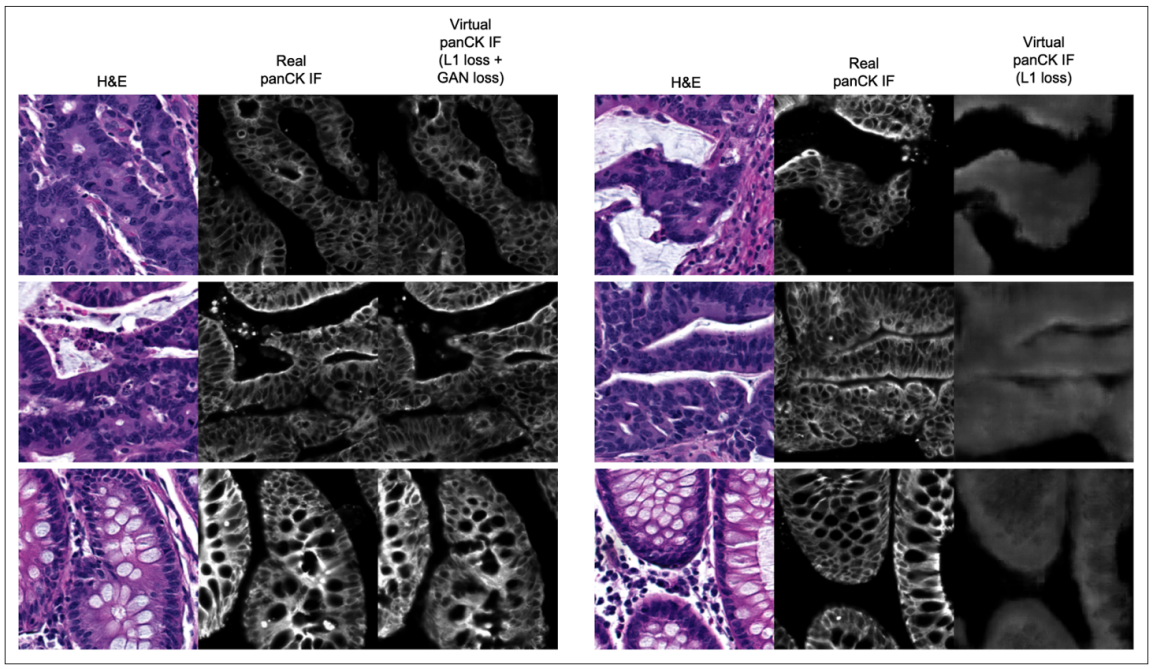


FIGURE 4.5: Left panels correspond to results from the full SHIFT model (generator and discriminator) and the right panels correspond to results from a model consisting of a generator only.

staining was to be possible using only H&E for prediction, it would be necessary for the H&E images to contain relevant biological information equivalent to that of CyCIF.

To test this hypothesis, I created tile-based image descriptors from H&E using a standard Variational Autoencoder (VAE)[57] and compared them to cell type composition vectors (7 cell types) created from CyCIF imaging data for the same tiles. In order to evaluate the overlap and exclusivity of each modality's information, I used canonical correlation analysis (CCA)[51] using two components. The two modalities quantitatively show high canonical correlations of 0.91 and 0.88 for each component respectively and qualitatively show a high level of overlap when the two components are plotted on top of one another (Figure 4.6A). Motivated by this

example, and building upon previous works in cross-domain data translation[72, 110], I built a cross-domain autoencoder (XAE) architecture which learns to co-embed H&E and CyCIF representations of the same tissue into the shared latent space (Figure 4.6B). To test a minimum working example of our XAE architecture, I performed a simple ablation experiment with the CyCIF encoder of the model removed. For this experiment, the model was tasked with H&E reconstruction and H&E-to-(DAPI and panCK) translation. To assess goodness of fit, the model was trained to convergence and evaluated on a validation set. Visual inspection of model outputs indicated that the model was functioning as intended (Figure 4.6C). In our original design, the XAE included skip connections that connected across the U-Net generator blocks, but I discovered that the models did not learn useful latent representations of images, a direct effect of the absence of loss function gradient flow through the interior layers of the models enabled by skip connections. I removed the skip connections in subsequent experiments and found that these models exhibit good convergence properties and have appreciable loss function gradient flow through the model interior (not shown).

Having confirmed that the trained XAE had fit its training distribution (Figure 4.6C), I next wanted to assess the representativeness and interpretability of the latent feature space that it learned with respect to pathologically interesting regions of the sample. To do this, I used the H&E encoder of the trained XAE to encode tiles from H&E test section 096 into 512-dimension feature representations and assessed how the features were distributed over tiles drawn from each of several pathologist-defined ROIs in the test section. The 6,742 non-overlapping tiles from H&E test section 096 which had at least one pixel

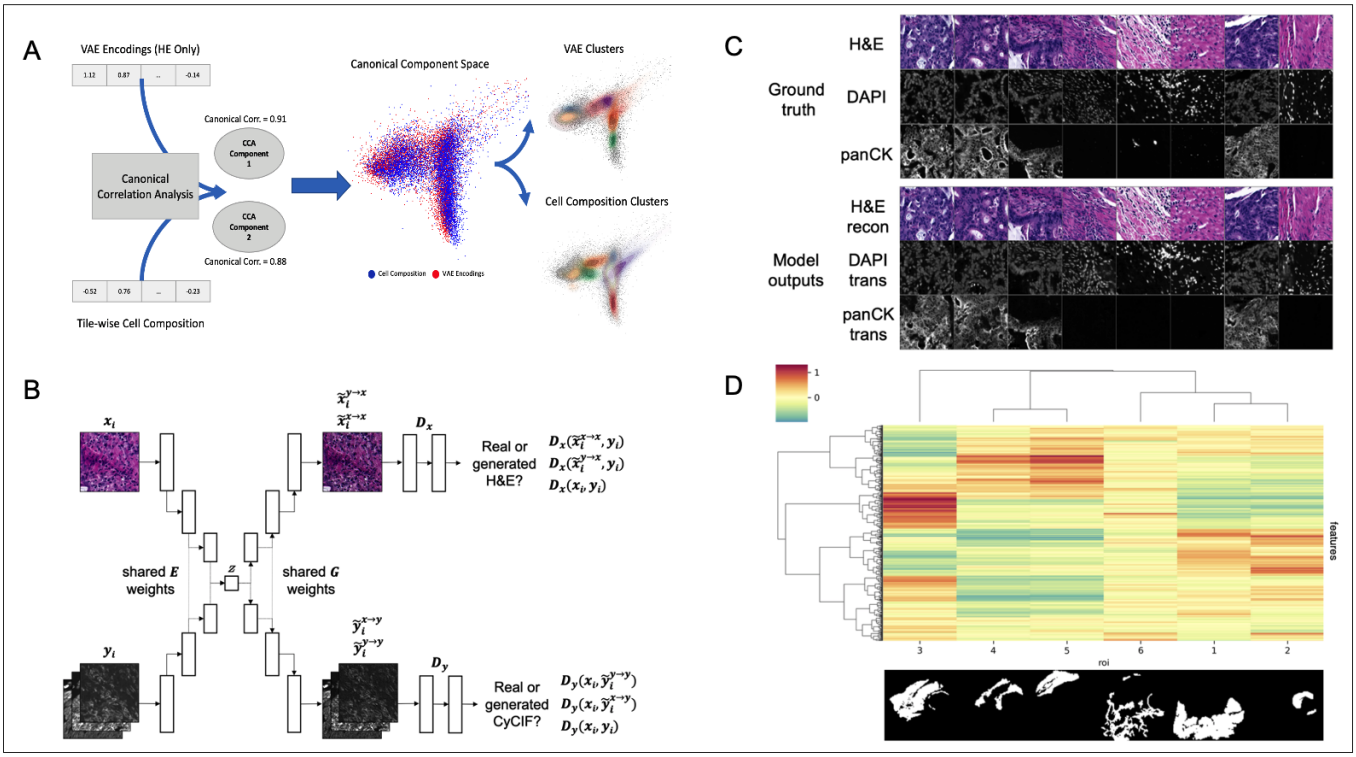
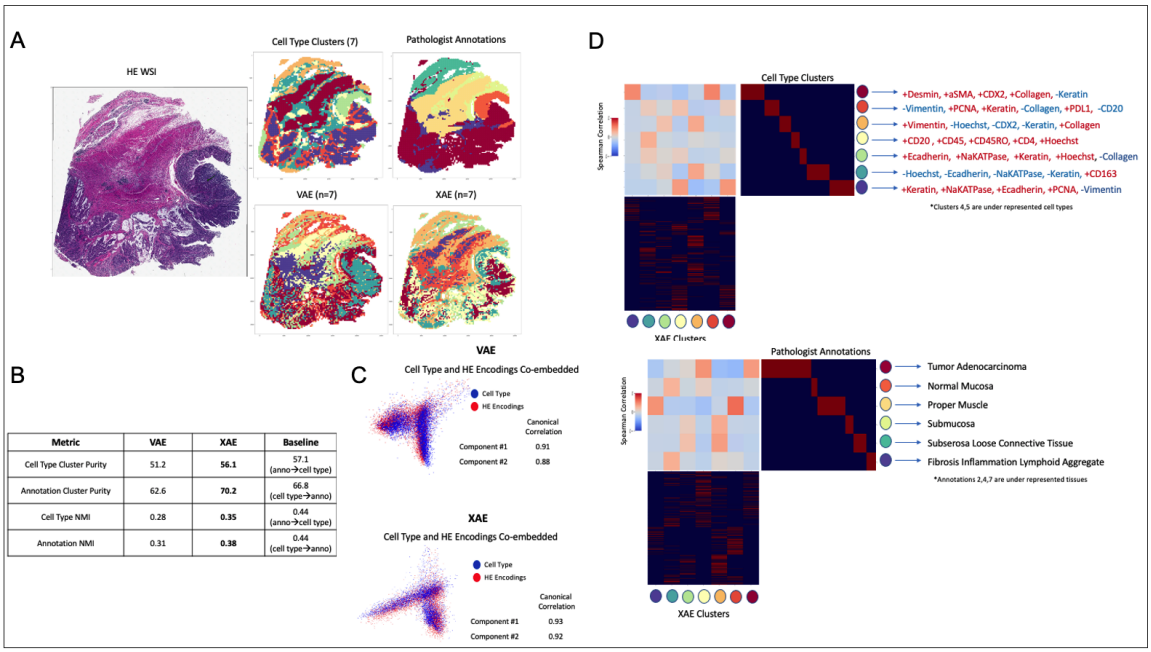


FIGURE 4.6: (A) VAE encodings of H&E and CyCIF cell type composition (7 cell types) show high canonical correlation and a large overlap between data and cluster embeddings. (B) XAE architecture. The model has two input heads, one for H&E encoder inputs (x_i) and another for CyCIF encoder inputs (y_i), both of which encode into a shared latent space (z).The model also has two output heads, one for H&E decoder outputs and another for CyCIF decoder outputs. Full XAE model architecture is described in Table 4.3. (C) Ground truth tiles representing a single training batch. Trained XAE model results for the tasks of H&E-to-H&E reconstruction and H&E-to-CyCIF translation using the ground truth training. (D) XAE latent feature clustering and corresponding pathologist annotation where the inset image indicates the binary mask corresponding to each ROI with respect to the layout of the H&E test section. Features were z-scored, then tiles were mean-aggregated based on their ROI and features were hierarchically clustered. The ROI label keys are 1: tumor adenocarcinoma (n = 2,501 tiles); 2: normal mucosa (n = 362 tiles); 3: proper muscle (n = 1,576 tiles); 4: submucosa (n = 473 tiles); 5: subserosa, loose connective tissue (n = 782 tiles); and 6: fibrosis, inflammation, lymphoid aggregate (n = 1,048 tiles). The color scale corresponds to the mean of z-scored feature values for each ROI.

of pathologist annotation were each encoded into 512-dimension latent feature maps. I found that many of the learned image features were associated with pathologically-distinct regions of the sample (Figure 4.6D).

In order to evaluate how well deep learning can capture and represent unseen complex information using H&E images alone, the VAE model features the XAE features (both generated from H&E images alone) were compared to cell types defined by CyCIF expressions and to pathologist tissue annotations. Clustering tiles within the WSI based on cell type composition using K-means resulted in 7 clusters, and the pathologist annotated 6 key tissue types to be used as ground truth (Figure 4.7A). Ground truth tile labels were compared against one another to create a baseline for evaluation. When annotations were used to predict cell type, there was a baseline performance of 57.1% cluster purity and 0.44 normalized mutual information (NMI). Conversely when cell type was used to predict annotations, there was a baseline performance of 66.8% cluster purity and 0.44 NMI (Figure 4.7B). In all metrics, XAE outperformed VAE predictions, achieving a 56.1% cluster purity and 0.35 NMI against cell type, and 70.2% cluster purity and 0.38 NMI against pathologist annotation (Figure 4.7B). It is also notable that on the metric of cluster purity against annotations, the XAE outperformed the baseline metric; this indicates that the XAE is better at predicting histologic tissue type than even cell type compositions.

Analysis of complex information, deeper than large scale clustering, was conducted using canonical correlations between the model embedding space and the tile-wise CyCIF expressions. Visually both VAE and XAE show a good overlap between cell type embeddings from CyCIF and model embeddings produced from H&E images (Figure 4.7C); the XAE, however, achieves higher canonical correlations (0.93 and 0.92 compared to 0.91 and 0.88 for VAE). To confirm that I was extracting relevant and rare cell types with the representation models, I computed



Metric	VAE	XAE	Baseline
Cell Type Cluster Purity	51.2	**56.1**	57.1 (anno→cell type)
Annotation Cluster Purity	62.6	**70.2**	66.8 (cell type→anno)
Cell Type NMI	0.28	**0.35**	0.44 (anno→cell type)
Annotation NMI	0.31	**0.38**	0.44 (cell type→anno)

FIGURE 4.7: A) Images colored by tile labels for cell type, pathologist annotation, assigned cluster from VAE using H&E input, and assigned cluster from XAE using H&E input. B) Quantitative evaluation of VAE and XAE at recapitulating biological labels, measured using cluster purity and NMI and compared to baseline of agreement between biological labels. C) Canonical correlation analysis between cell type composition vector and H&E encodings for both VAE and XAE, quantitatively measured by component correlation and qualitatively by label overlap in embedding space. D) Cluster-wise correlation matrix for XAE against both cell type and pathologist annotations to determine which biological features are adequately captured. Defining CyCIF expressions provided based on inter/intra-cluster variability.

the Spearman correlation between every predicted cluster and ground truth cluster (Figure 4.7D). From this I can see that XAE has consistently high magnitudes of correlation, and that a reasonable correlation exists for every ground truth cluster except for cell type clusters 4 and 5 which are underrepresented populations. Furthermore, the cell types that the XAE is able to capture are largely explained by changes in Na-K ATPase, E-Cadherin, and PCNA, which were shown to be important indicators for cell phenotypes in prior research on this tissue[67].

It is shown by several metrics that the XAE model outperforms the VAE in capturing detailed information from H&E images alone, which are able to adequately recapitulate information from CyCIF expression data and pathologist annotations that are unseen during test time. Because the XAE encodings are able to adequately recapitulate the information in CyCIF from H&E, I can use them for proxy analyses such as selecting representative regions of the WSI to be stained or analyzed with other modalities and methods.

B. Co-embedding H&E and IF representations improves ROI selection

Currently ROI selection within WSIs is done either randomly, which is inaccurate and is likely to select an area that doesn't represent the WSI, or with manually, which is biased, subjective, and has been shown to miss whole tissue patterns[67]. The XAE encodings described above capture the complex cell type and annotation information using H&E alone, which means that the information it extracts can be used to quantitatively evaluate the features within regions across imaging domains. Using these XAE extracted features, I develop an optimization-method to select a minimum set of ROIs that are more representative of the whole slide features than random sampling. The additional benefit of this approach is that it is repeatable and biologically-driven, so multiple people and labs can perform the same analysis with the same results. To evaluate ROI selection performance, I used three metrics: mean squared error (MSE) between the cell type composition of selected ROIs and WSI; Jensen-Shannon Divergence (JSD) between the cell type composition vectors of selected ROIs and WSI; and mean entropy of the selected ROIs' cell type compositions. Since MSE and JSD both operate on different

principles, quantifying individual values and overall distributions, the use of both for evaluating composition is beneficial. While MSE is highly prone to outliers and abnormal data, amplifying error of single erroneous samples, JSD provides a smoothed and normalized metric. Three different methods for ROI selection were tested: random sampling, convex optimization minimizing l_1-norm of cell type composition, and convex optimization minimizing l_1-norm of cell type composition with maximizing entropy to select ROIs with more heterogeneous cell composition.

When regions are randomly sampled, one can see that the cell type compositions struggle to converge to the whole slide cell type composition, taking upwards of 20-30 ROIs (each of which comprises between $\sim$0.15% and $\sim$0.80% of WSI area individually) before reaching a reasonable representation (Figure 4.8top). Using a simple composition-based optimization drastically decreases the number of ROIs necessary to $\sim$7. This number of ROI is equivalent to the number of cell type clusters I was optimizing for and further investigation shows that the algorithm was selecting primarily homogeneous regions that reconstruct the whole slide composition. This is validated looking at the mean entropy of ROIs for the base convex optimization method, which consistently shows low to middling ROI entropy values, especially at the 1000 pixel size data where there is decreased chance of getting diverse populations simply due to the ROI size (Figure 4.8middle).

To select a more heterogeneous region, entropy is considered in the convex optimization, and convergence is observed much earlier at 3-4 representative ROIs (Figure 4.8top). Unlike the simple optimization considering cell composition only,

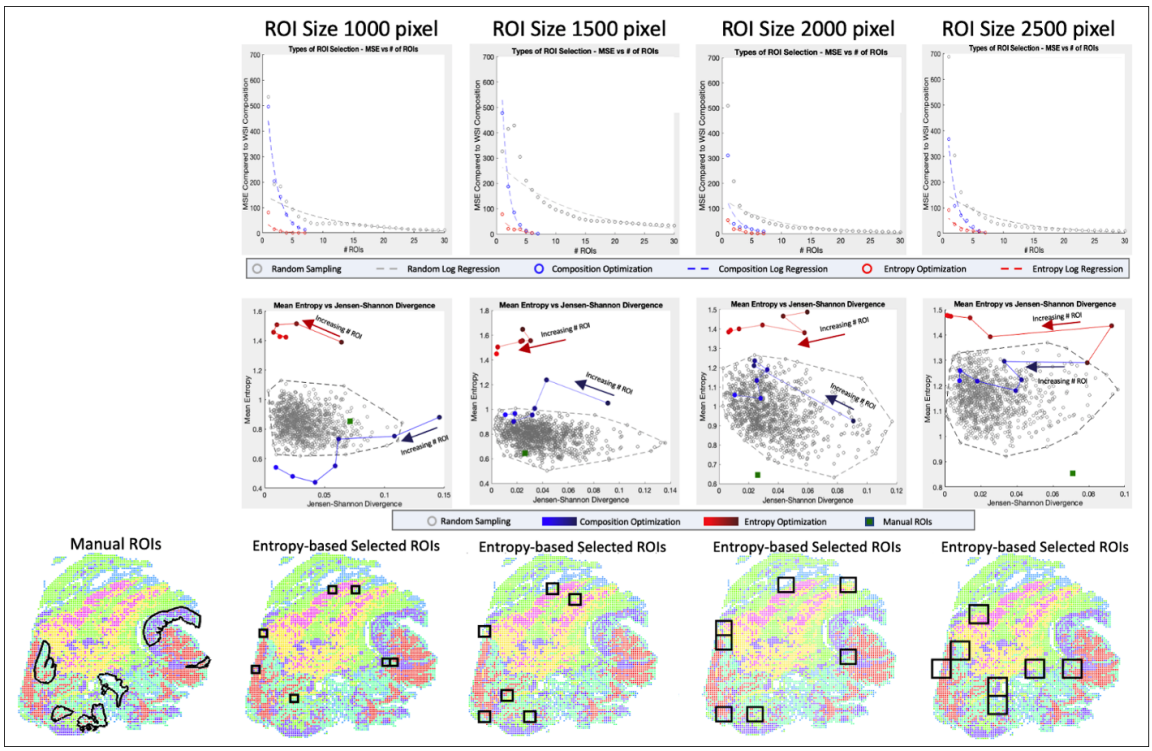


FIGURE 4.8: For four ROI sizes (1000x1000, 1500x1500, 2000x2000, 2500x2500 pixels) and three sampling techniques (random sampling, convex optimization using cell type composition, convex optimization using cell type composition and regional entropy), I calculate the optimal selection of ROI. Top row) By calculating the MSE for a range of ROI, I can evaluate the rate and quality of convergence for each technique. Middle row) Selections of representative ROIs are evaluated based on two metrics (Entropy for tissue heterogeneity and Jensen-Shannon Divergence for composition similarity.) Random sets of 7 ROIs are generated 1000 times to portray the baseline pattern. Selections from linear and convex optimizations are plotted with increasing numbers of ROIs to show the change in performance. The performance of the manually selected ROIs is also shown to emphasize the bias in targeted sampling. Bottom) The optimal ROIs are shown for convex entropy optimization at each size of ROI. Image colors portray the XAE labeled clusters describing cell and tissue type.

however, the ROIs selected are not homogenous and include much more biologically interesting regions with diverse cell populations. This is confirmed with entropy values considerably higher than the randomly sampled population (Figure 4.8middle). When looking at the full range of clusters, both optimization-based approaches are substantially better than even manual ROI selection which is

extremely biased, scoring poorly on both composition metrics and heterogeneity metrics.

Manual annotation is often guided by a desire to sample a specific set of tissue types, in this focusing in on three tissue types (tumor adenocarcinoma, normal mucosa, and lymphoid aggregate) while being unable to account for cell type. To account for this and provide a more fair comparison, I narrowed the range of clusters being optimized for in the ROI selection to only consider tiles with the relevant cluster identities (Figure 4.9). Even in this restricted cluster set, manual annotation is less representative of the WSI's tumor and immune cell type composition as produces less heterogeneous regions. This shows that the improvements made over manual selection are not solely due to the tissue type bias of pathologists selecting interesting regions; it is also the fact that the ROI selection based on convex optimization method can find the most representative regions which can be a difficult task for an annotator who cannot see cell type without substantial time and effort.

4.3.3 Optimized panel selection to maximize marker predictability

A. Proof-of-concept for using a generative model to impute missing markers

As I have shown in the above sections, deep learning enables the use of information from one subset of data to predict information about another subset when there is a significant amount of mutual information between the two. Just as it is possible to predict the spatial staining patterns and intensities of CyCIF from H&E, it is also possible to predict the CyCIF staining of one panel using another. This

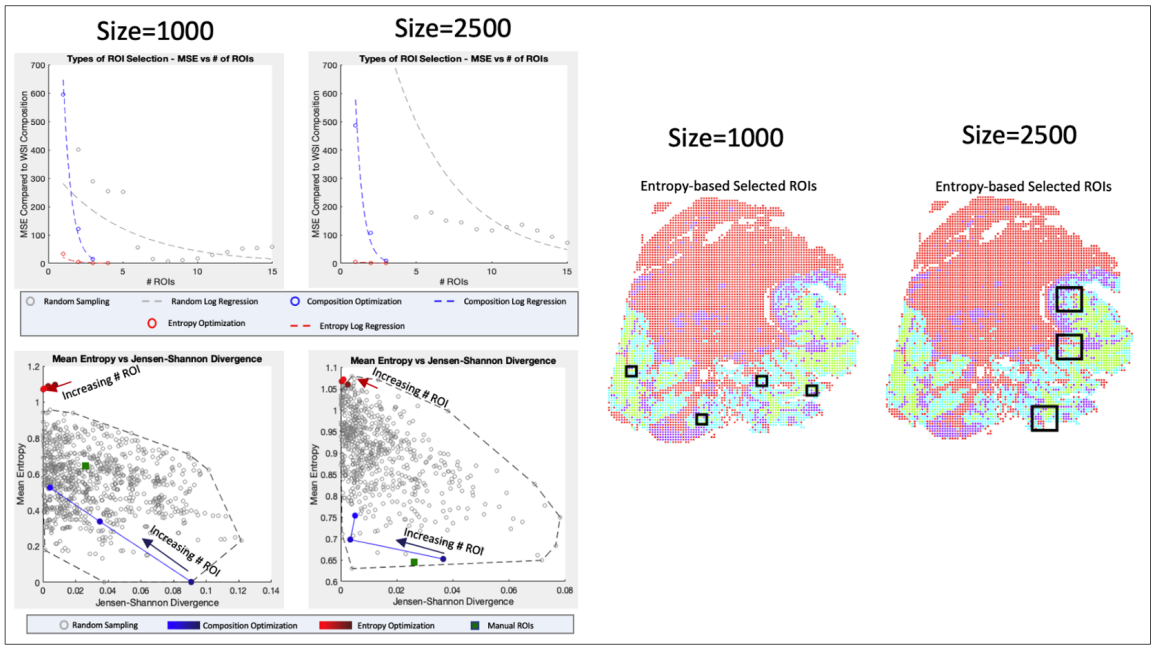


FIGURE 4.9: For two ROI sizes (1000 pixel, 2500 pixel) and three sampling techniques (random sampling, convex optimization using cell type composition, convex optimization using cell type composition and regional entropy), I calculate the optimal selection of ROI. Prior to calculation, I restrict the cluster type identities being optimized for to only those that were targeted by manual annotation. Top row) By calculating the MSE for a range of ROI, I can evaluate the rate and quality of convergence for each technique. Middle row) Selections of representative ROIs are evaluated based on two metrics (Entropy for tissue heterogeneity and Jensen-Shannon Divergence for composition similarity.) Random sets of 7 ROIs are generated 1000 times to portray the baseline pattern. Selections from convex optimizations are plotted with increasing numbers of ROIs to show the change in performance. The performance of the manually selected ROIs is also shown to emphasize the bias in targeted sampling. Bottom) The optimal ROIs are shown for convex entropy optimization at each size of ROI. Image colors portray the XAE labeled clusters types containing information on both cell type and tissue type (red being cell tiles not considered in this analysis).

can be used to reduce the number of stains in CyCIF protocols by using a reduced panel set to predict a larger panel of markers without actually having to stain for them. The question then becomes, what is the theoretically best selection of markers for maximizing the amount of information retained and generating the whole panel image predictions.

To test this process, I used a breast cancer TMA dataset comprised of 88 cores, 6 different breast cancer subtypes (plus normal), and 25 markers in the CyCIF tumor panel (Table 4.1) as described in the section 4.5. I evaluated 4 different methods for selecting an optimal reduced panel set (random selection, correlation-based selection, gradient-based selection, and sparse subspace-based selection), as shown in Figure 4.10 and described in section 4.5. The reduced panels of every method were then used to reconstruct the initial full panel using a variational autoencoder (VAE), which encodes the markers in the reduced panel into a latent descriptor and generates all 25 markers in the initial panel set. The reconstructed images of each method are then evaluated using two common analytics (mean intensity correlation and cluster overlap) to determine whether information is retained in the reduced panel and prediction pipeline.

Channel	Marker		Channel	Marker
1	DAPI		14	Ki67
2	CD3		15	CD45
3	ERK-1		16	p21
4	hRAD51		17	CK14
5	CyclinD1		18	CK19
6	VIM		19	CK17
7	aSMA		20	LaminABC
8	ECad		21	Androgen Receptor
9	ER		22	Histone H2AX
10	PR		23	PCNA
11	EGFR		24	PanCK
12	Rb		25	CD31
13	HER2		-	-

TABLE 4.1: Full TMA panel set

As a proof of concept to further demonstrate how information from a reduced set can adequately predict unseen information within the full set, I randomly selected

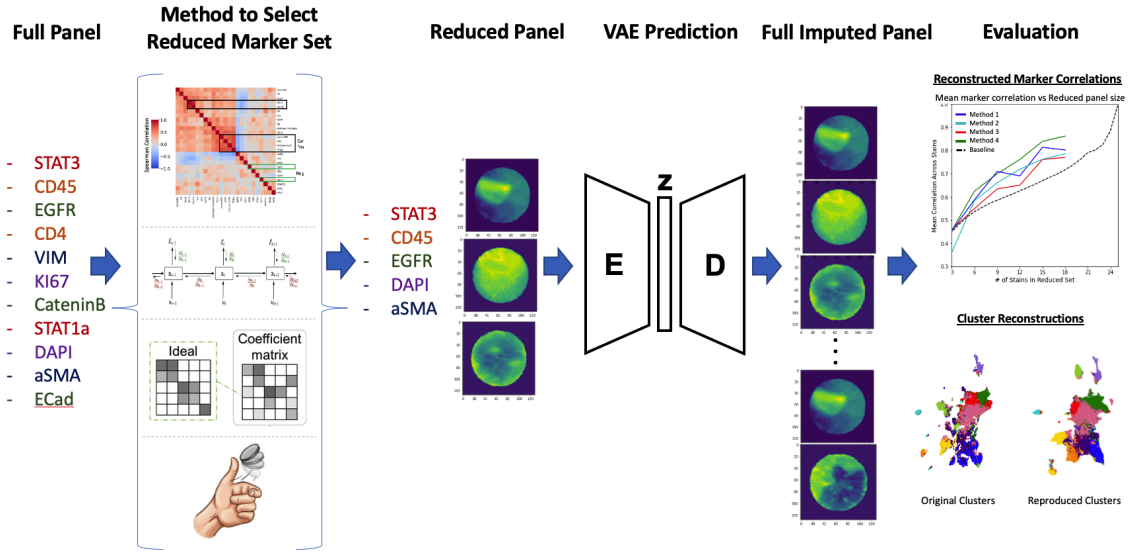


FIGURE 4.10: In order to select an optimal reduced panel from a designed full panel, four different selection methods were tested: intensity correlation-based, deep learning gradient-based, deep learning sparse subspace-based, and random selection. Using the reduced panels selected from each method, a VAE was used to impute the full panel set. The full set of imputations were then evaluated by comparing them to the original images using expression correlations and cluster overlap as these are two important features of downstream analytics.

50% of the full panel (Table 4.7) which was used to predict the other 50% (Figure 4.11). As can be seen qualitatively in the real and predicted image pairs, the morpho-spatial features of size, shape, distribution, and relative intensity are preserved, regardless of whether the marker was present in the training panel. Quantitatively this is captured using the structural similarity index measure (SSIM), which measures the perceptual difference between two images. The overall quantification shows a mean SSIM of 0.75. The predictions also achieve Spearman correlations of 0.91 and 0.75 for included stains and withheld stains, respectively.

To help ground the quantitative metrics in this proof-of-concept experiment in a

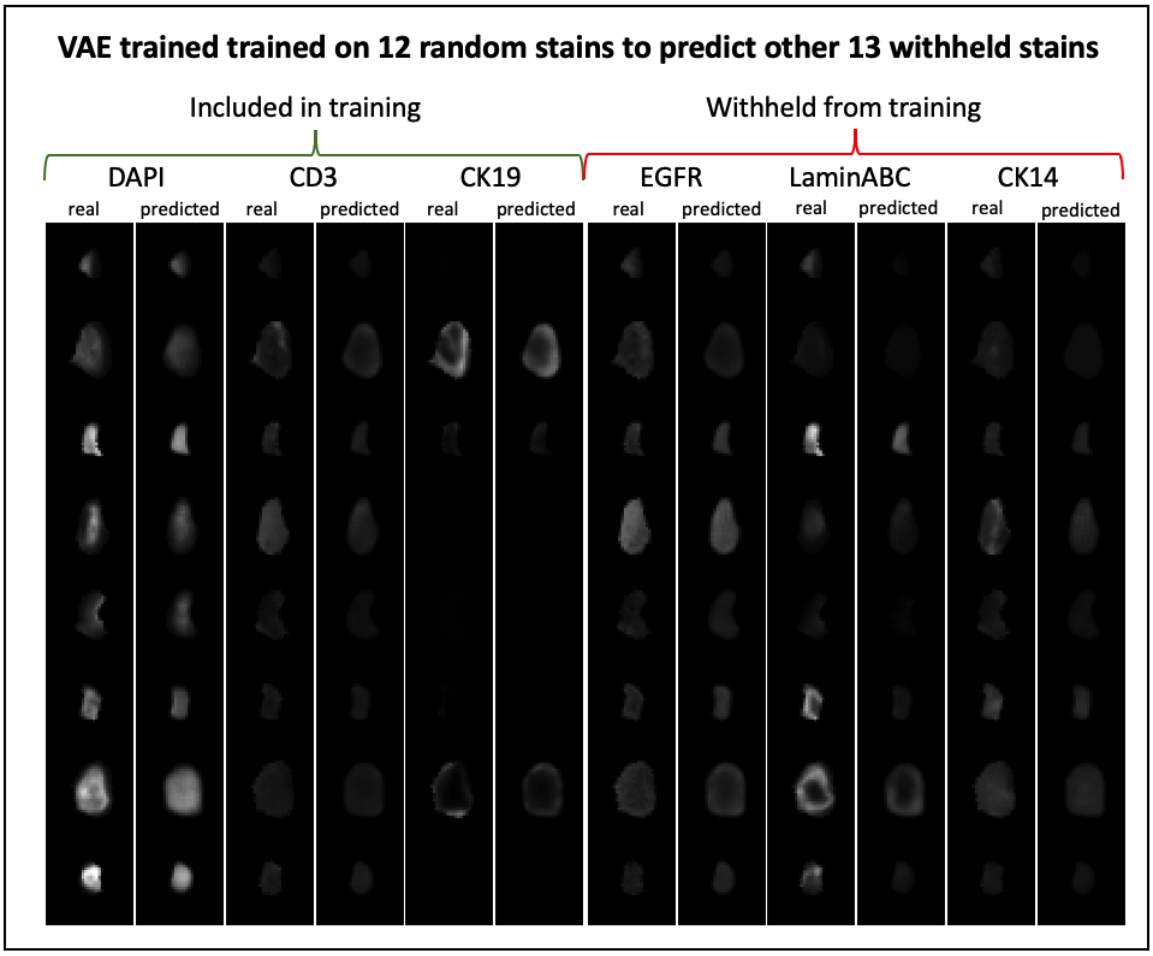


FIGURE 4.11: 12 stains were randomly selected to create a reduced panel which were then used to train a VAE to reconstruct the full panel of stains. Here I show a representative 3 stains from the included and withheld marker sets across 8 cells. Real and predicted staining is shown side by side to qualitatively demonstrate that a reduced panel can reconstruct relevant unseen information.

more interpretable context, I compare the degree of error to other forms of common technical noise (blurring, salt/pepper, and differences in segmentation as simulated by erosion/dilation) as can be seen in Figure 4.12. One can see visually and quantitatively that each of the different noises have varying degrees of

severity, with blurring having the smallest effect on intensity and overall structure of the image. Differences in segmentation via erosion/dilation has the largest effect as the inclusion and exclusion of a few pixels can make a significant difference when the overall size of the cell is only about 10-20 pixels across[119]. The effect of this mis-segmentation will also have a larger effect on membrane stains or densely packed cell populations. With regards to the structure of the predicted image, the randomly selected panel of 12 stains achieves an average SSIM of 0.75, just behind blurring at 0.78 and well above salt/pepper and erosion/dilation at 0.68 and 0.67, respectively. Although the predictions from the randomly chosen panel of 12 stains rank lowest in mean intensity correlation at 0.80, the score is still comparable to segmentation noise at 0.83. This shows that a variational autoencoder can recapitulate the full panel image using a randomly selected panel that is half the size of the original, but does so with differences that are similar to the variation created by normal technical imaging artifacts. Evaluation of the utility of these reconstructions will be discussed in later experiments. The question still remains, however, to what degree selecting the reduced panel methodologically can improve these results.

B. Evaluating selection methods with imputed marker correlation

Our first evaluation of panel reduction methods was conducted by correlating the original and reconstructed panels' mean intensities (Figure 4.13). This was done over an increasing number of stains in the reduced set to show how each method performs with different levels of information available in the reduced set. For a baseline comparison, the intensities of the reduced panel were used as a 1-to-1

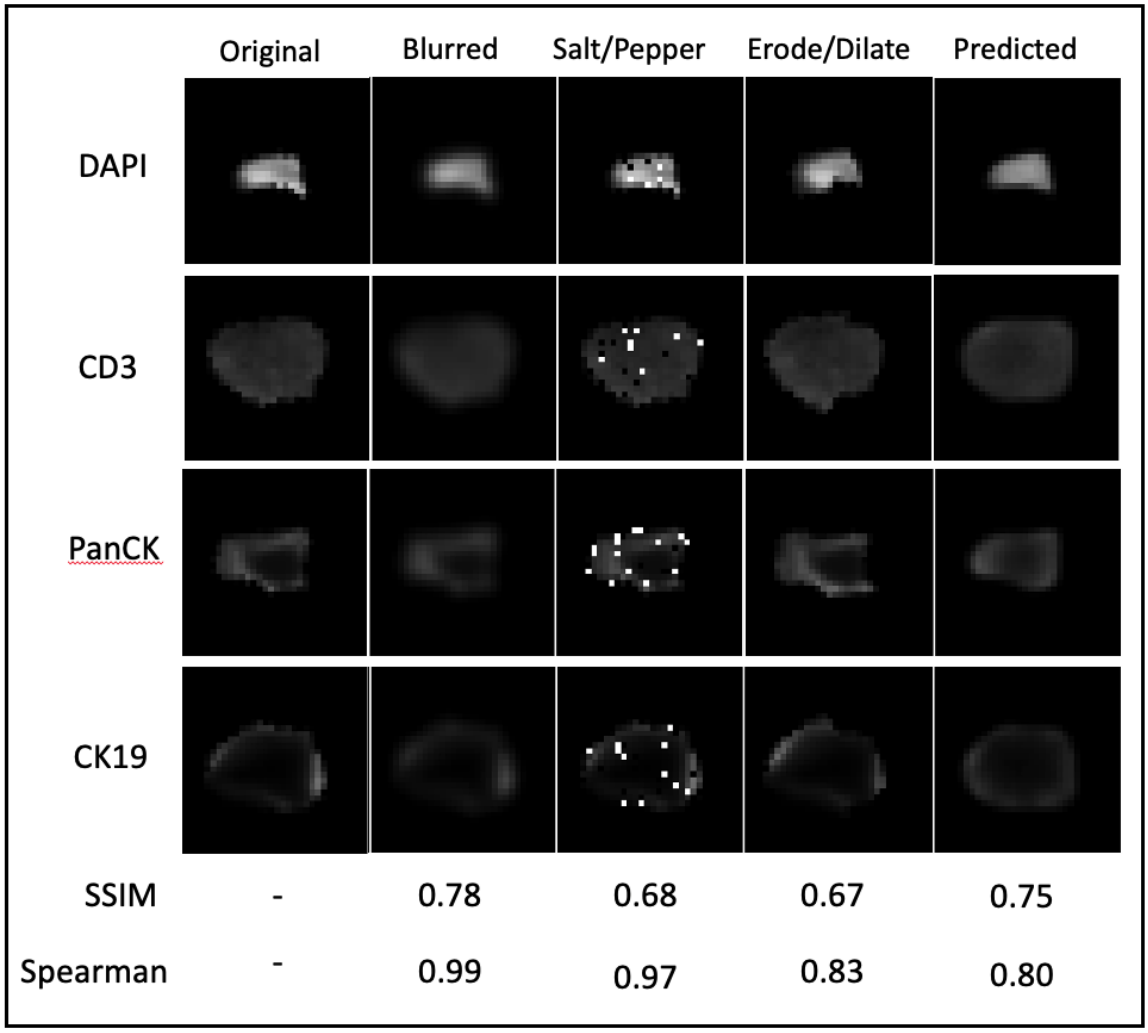


FIGURE 4.12: To frame the extent of error in the predicted results from a randomly selected reduced panel of 12 stains, several technical noises were simulated and evaluated for the same metrics. The average SSIM was measured for each stain individually and averaged. Likewise, the Spearman correlation between the original stain intensity and the resultant stain intensity was calculated for each stain independently and averaged across the withheld panel set.

substitute for the missing stains; for example, if CK19 is included in the reduced panel and PanCK is not included, then the CK19 expression will be directly used as the prediction of PanCK expression since it is PanCK's highest correlate within the reduced panel. This baseline of 1-to-1 substitution resulted in subpar predictions

that did not converge to a correlation approximating the full panel until nearly all stains were included in the reduced set, indicating the need for predictive models to retain information on removed markers. Random selection performed moderately better than baseline, achieving a mean Spearman correlation of 0.77 for withheld markers and 0.89 for all markers when 18 of 25 markers are included in training. The reasonable performance of the random sampling method, both here and in the above proof-of-concept, is primarily a result of how well deep learning models can process and predict patterns missing from the data. Random selection here can be used as a computational baseline to illustrate the increased performance and predictive power that comes from deep learning regardless of intelligent panel design. Without the constraints of time and processing power, this study could be improved by training and evaluating predictive models using dozens of randomly selected panels; however, each deep learning model (for each panel arrangement and size) can take more than a day to train and evaluate. For this study only a single randomly selected panel at 6 different sizes is used, which prohibits analysis as to the variability of random predictions.

Sparse subspace selection performs slightly better than random selection and baseline, achieving mean Spearman correlations of 0.80 and 0.89 for withheld markers and all markers, respectively. By looking at the correlations with respect to panel size (Figure 4.13), however, one can see that subspace-based selection performs even better at lower panels sizes compared to random. Gradient-based selection performs better than random or subspace-based selection methods within the withheld marker predictions, achieving a max correlation of 0.81. It is worth noting

that this prediction method appears to be less stable with prediction metrics fluctuating when different panel sizes are used. Within the full marker set, gradient-based selection performs similarly to random and subspace selection methods, achieving a correlation of 0.88. Finally, correlation-based selection also performs well at reconstructing the mean intensities of each stain, both within and withheld from the reduced panel. For most every panel size, correlation-based selection achieves the highest Spearman correlations compared to the other selection methods, obtaining a correlation of 0.86 and 0.90 for withheld and all markers, respectively. For the purpose of reconstructing mean intensity information, the other selection methods only perform similarly to correlation-based selection at extremely low panel sizes where there is insufficient information.

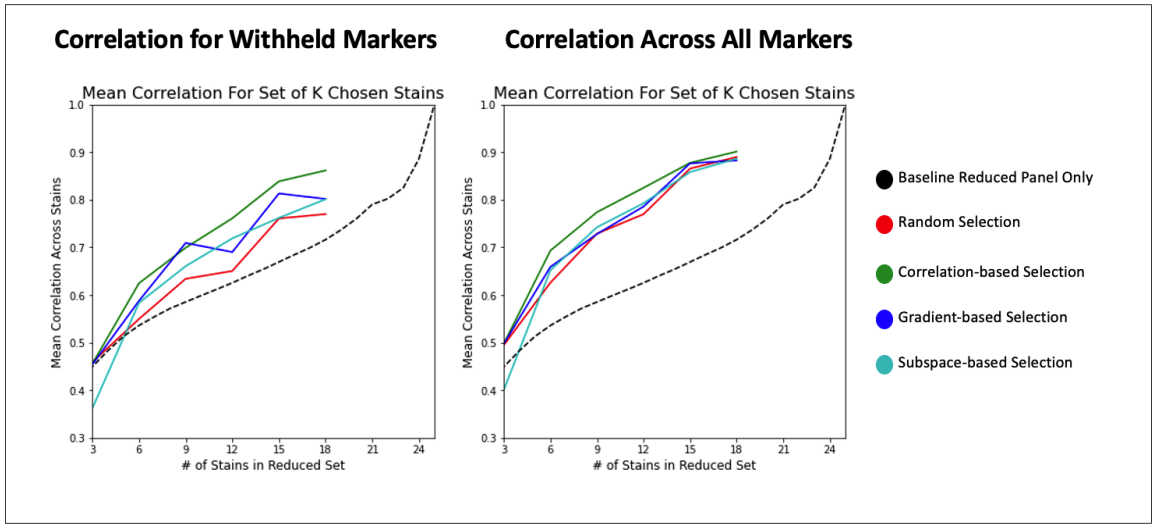


FIGURE 4.13: All panel selection methods were evaluated across a range of panel sizes to determine how well their reduced panels can be used to reconstruct the full panel. Spearman correlation was measured for each stain independently and then averaged across the whole dataset. The data was split into withheld markers (left) and all markers (right) to illustrate each model's generalizability and performance in both domains. 1-to-1 substitutions of marker intensity were used as the baseline, where makers withheld from the reduced panel set were simply assigned the intensity of their closest match as described in section 4.5

In order for a panel to be consistently effective, it must be generalizable across datasets and similar pathological states. An example of this is breast cancer subtyping where unique expression patterns will vary and classification can fail if important markers are not able to be predicted properly based on the underlying information of that specific biology. To test this generalizability, I applied the highest performing panel (correlation-based with 18 markers) to all the different cancer subtypes within the TMA dataset separately and evaluated the predicted expressions (Figure 4.14). Although there is some slight variance, the panel performs well consistently across all subtypes, achieving Spearman correlations between 0.72 and 0.83 within the excluded markers. However, the specific markers that scored the highest and lowest correlations in each subtype, did vary based on the relative biological expression. It can be further observed in Figure 4.15 that the markers that performed poorly simply did not have large variation across the specific subtype. This can be seen distinctly in PR, H2AX, and PCNA. The predictions receive poor correlations for all subtypes when the marker does not show substantial positive expression, and the predictions receive good correlations whenever the subtype does show a variable expression range. Although many of the low correlation scores around 0.70 are still adequate, their reduced performance compared to the other markers is due to their low variability in a cancer subtype. This can be further seen in PCNA where the correlation metric is 0.39 when the marker is completely absent from the subtype. This absence of markers skews the evaluation of the models. This also further illustrates the consistency of the panel across subtypes despite the differences in biology and marker expression, since the predictions only score low correlations for absent and low variability samples.

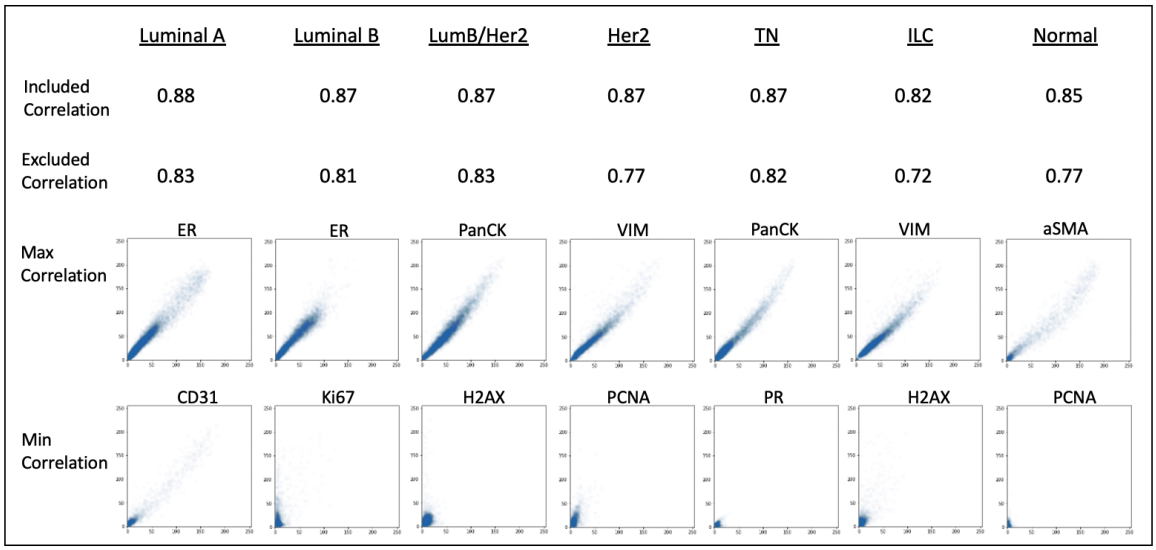


FIGURE 4.14: The full panels of six different cancer subtypes and normal were predicted using the highest performing reduced panel (correlation-based selection with 18 markers). Spearman correlations were calculated between the full panel expressions and the expressions of the predicted markers for the included markers and excluded markers separately. The expression correlation plots for the best and worst predicted markers are shown for each subtype.

C. Evaluating selection methods with cluster matching

Although it is important to be able to reconstruct the mean intensities of cells, downstream analysis such as single cell phenotyping and clustering is important for biological research and if such analytical methods were to be affected, then the reduced panel predictions would not be useful for complex research methods. As shown in Figure 4.16, although the selection methods have varied levels of performance at predicting mean intensity, when 18 of 25 markers are included in the reduced panel sets, all selection methods perform well at recapturing the same clusters extracted from the full panel set, as measured by normalized mutual

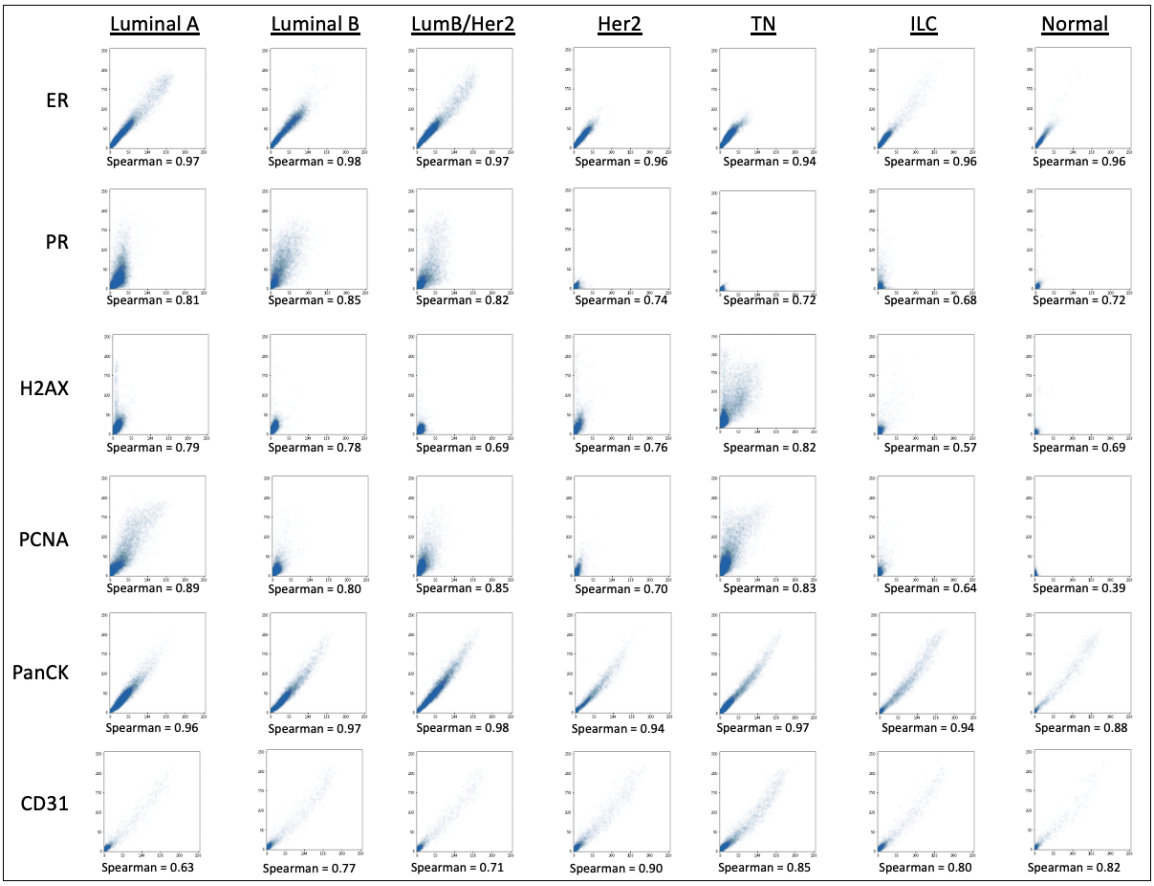


FIGURE 4.15: A sample of a few of the lowest scoring and highest scoring makers were selected to directly compare the Spearman correlations across all the breast cancer subtypes. Predicted and true expression were compared for each marker and subtype individually.

information (NMI)[96]:

$$NMI(U,V) = \frac{MI(U,V)}{mean(H(U),H(V))} \tag{4.1}$$

where U and V are the reduced panel predicted and full panel (ground truth) cluster labels and H(U) and H(V) represent the entropy of U and V, respectively. The predicted clusters were then paired to their full panel counterpart by examining the population compositions to maximize consistency.

This again shows that the information within the 7 withheld markers is able to be predicted using the 18 markers in the reduced panel, enough to produce similar downstream results for clustering and potentially cell-type calling. Although it would be ideal to compare results to ground truth cell types, the dataset was limited by lack of labeled data; therefore, the clustering results from the full panel mean intensity dataset were used as ground truth for the single cell populations that can be extracted. The correlation-based method achieved the highest NMI of 0.64, while gradient-based, subspace-based, and random selection achieved only slightly lower NMIs of 0.60, 0.63, and 0.60, respectively. All clustering results are significantly larger than the baseline NMI of randomly shuffled cluster labels (NMI $= 0.0002$). Although the spatial distance in the illustrated UMAP cluster plots is not quantitative in regards to similarity of clusters, one can qualitatively see the same overall pattern and organization of clusters between the full panel and all selection methods. This shows that while intelligent selection of the reduced panel will matter for the overall prediction of the expression information, the selection method might be irrelevant to the end result of other analytics such as clustering. This is because deep learning architectures can learn to capture the most defining information so long as they are trained on large enough panel size, regardless of which channels are used in the reduced panel.

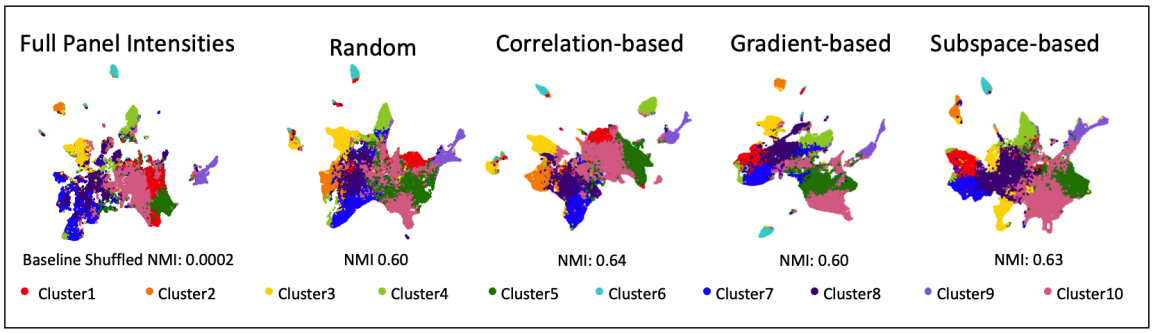


FIGURE 4.16: Clustering was performed on the full panel intensities to generate ground truth cell type clusters using k-means (k=10, chosen with elbow method on silhouette score). Random, correlation-based, gradient-based, and subspace-based selection methods were also clustered using reconstructed intensities as input to k-means (k=10). Clustering similarity to ground truth was performed using normalized mutual information (NMI). A baseline NMI for comparison was generated using randomly shuffled cluster labels. The clusters were projected into a UMAP embedding and plotted to visually show the cluster results. Colored cluster-labels for all prediction methods were applied by matching cell compositions with the full panel.

4.4 Discussion

Tumors are heterogeneous tissue volumes up to centimeters in size that can hardly be represented by a microscopic 2D section, but many of the imaging characterization platforms in both research and clinical practice make the assumption that TMAs containing small core samples of essentially 2D tissue sections are a reasonable approximation of bulk tumor. However, emerging 3D tumor atlases strongly challenge this assumption[67, 56, 29]. In spite of the additional insight gathered by measuring the tumor microenvironment in 3D, it can be prohibitively expensive and time consuming to process tens or hundreds of tissue sections with CyCIF. Even when resources or time are not limiting, the criteria for ROI selection in tissues for downstream analysis remain largely qualitative and subjective. Furthermore, excessively large and self-correlated staining panels increase the amount of time, effort, and expense needed to obtain images and can result in decreased

image quality in later rounds of imaging.

In the current study, the virtual staining paradigm is extended to a 3D CRC atlas[67] and demonstrate a proof-of-concept that generative models can learn from a minimal subset of the atlas to reconstruct the remaining sections of the CyCIF portion of the atlas and recapitulate quantitative endpoints derived using the real CyCIF data. Quantitative comparisons of real and virtual CyCIF stains exposed the challenge of using adjacent sections to train models, where image contents are subtly but appreciably different between sections at single cell resolution. This challenge could be overcome in future studies by staining each tissue section first with CyCIF, then terminally with H&E[133]. That being said, this study and those like it take for granted that histology workflows are inherently destructive, since serial sectioning and processing of tissue can preclude tissue from being used in other assays. Alternatively, a non-destructive 3D microscopy approach using tissue clearing and light-sheet microscopy could be deployed, which would also preserve tissues for other assays[71]. However, the slow diffusion rate of antibodies in whole tissues limits the deep multiplexing potential of the CyCIF platform in this non-destructive approach, but the use of small molecule dyes and affinity agents could help to overcome this challenge to 3D virtual staining applications[152].

I also implement and evaluate a novel deep learning model which integrates paired H&E and CyCIF data into a shared representation and demonstrate that the model can be used as a quantitative and objective guide for ROI selection, with the integrated H&E/CyCIF representations being more informative than H&E representations alone. The limitation of this approach is that the XAE model must be trained using paired H&E-CyCIF data prior to being used for prediction and

quantification. A further limitation is that the ROI selection can only be optimized with respect characteristics that can be quantified, such as heterogeneity and composition used here.

Although image representations can accurately describe biological features, they cannot convey what may or may not be biologically interesting to researchers or clinicians. Cell type composition and entropy were used as metrics of biological relevance in this setting, but it is likely that other experiments will have different priorities. Some examples of this might include: weighting cell type clusters by level of interest; weighting entropy negatively if homogeneous regions are desired; weighting some other extracted metric such as co-localization of two cell types of interest. If one, for example, wishes to select the minimum number of ROIs that capture all the potential tumor cell states, then one would simply need to add a metric quantifying whether the cell types are included to the objective function, assuming that the encoded latent space created by the XAE were successful at capturing all the tumor cell types. Ultimately, the proposed method of optimization is versatile and amenable to many different functions depending on the biology and the needs of the researcher. The key takeaway is that this pipeline allows for intelligent representation from H&E images, which enables a plethora of subsequent analyses on this representation space with other multiplexed imaging platforms such as MIBI, IMC or GeoMX when only a few ROIs can be selected and analyzed using these platforms. The current methodology requires training the XAE model on a subset of paired CyCIF and H&E images, limiting its broad application unless a model is trained to be generalizable. The method, however, still has great potential to save time and resources as it can be applied to parallel slides and 3D

volumes to select regions for analysis throughout.

Finally, I evaluated several methods for optimally removing self-correlating markers from existing multiplex panel sets, while retaining the maximum amount of information by using generative deep learning to predict the staining of the full image panel, including withheld markers. The limitation of all these methods of panel selection is that they require a round of staining to be conducted first so that the marker interactions can be measured and evaluated to determine the level of panel self-correlation. Ideally, it would be best to use the information gained from these selection methods to design panels for new datasets without having to stain, test, and design for every new dataset, tissue, or patient. Although here I show the method's utility for identifying reducible markers within a single diverse dataset of multiple cancer subtypes, future research can look into the deployment of the designed panels to new datasets without the need for retraining. Research can also be done into the biological relevance of the reduced sets so that researchers can better design panels on their own with fewer excess predictable markers. By identifying which markers are consistently well predicted and which consistently fail regardless of panel reduction method, researchers can design future panels with informed decisions to include the poorly performing markers and exclude the easily predicted markers. It is worth noting, however, that the reduction methods used here only look to remove markers on the basis of expression and information redundancy, by making a panel that can repredict the full panel. There are many more factors to consider when designing a panel, such as wavelength overlap, competition between markers, and duplicate markers used for quality control. The methods proposed here can help guide the design of panels, but in no way are

fully able to replace the thought and decisions needed in many other aspects of the experiment.

In conclusion, all three tasks performed in this chapter (3D stain propagation, guided ROI selection, and intelligent panel reduction) demonstrate the capacity for deep learning to lessen the burden on researchers and potentially guide future experiments. These techniques are able to decrease the time and cost of analyses, while also making quantitative decisions that might otherwise be subject to researcher bias. Deployment of these methods will benefit multiplex staining pipelines that are currently bogged down in tedious redundancy and large datasets.

4.5 Methods

4.5.1 3D stain propagation methods

A. H&E and CyCIF image normalization

It is extremely common for stain distributions, intensities, and colors to vary between images, even for common staining method like H&E and even when the images were acquired within the same lab. These staining variations can make it difficult for deep learning models to generalize to images that look different from the images they were trained on. To minimize the influence of technical variability on stain color between H&E sections, the application of several stain normalization

methods to the H&E WSIs[98, 134, 76] were tested using the Python package stain-tools[1]. To identify and mask out background regions of each WSI (white regions of slide without tissue), WSIs were each cropped into non-overlapping 256x256-pixel tiles and tiles containing greater than 70% area of pixels with 8-bit encoded intensity greater than $(210, 210, 210)$ were excluded from subsequent normalization steps. To help identify and mask out background pixels in the cropped and accepted foreground before model fitting and normalization, the foreground tiles from each H&E WSI were independently standardized such that 5% of all pixels were luminosity saturated. For all normalization methods, the H&E WSI from middle most section is used as the stain reference to which the stain intensity distributions of all other H&E WSIs would be fit. Foreground tiles of each non-reference WSI were used as a whole to determine the color distributions of the stain in that section, and all foreground tiles for each section were then normalized at one time to fit the reference distribution. After the foreground tiles were normalized, they were restitched to form cohesive WSIs. On the basis of visual inspection (Figure 4.2B/C), I opted to use the Reinhard normalization method, which has also been shown to maximize deep learning model performance on digital pathology applications[128]. To control for variations in raw contrast between CyCIF WSIs, the intensities of CyCIF WSIs were re-scaled to have a min-max range fit to the 70^{th}-99.99^{th} intensity percentiles of the input WSIs.

[1]https://github.com/Peter554/StainTools

B. 3D registration of paired H&E and CyCIF

The construction of a 3-dimensional volume requires that all the points are spatially aligned and continuous throughout all axes; otherwise, the resulting volume would be disjointed with erroneous fragments, projections, and holes. Because image layers within 3D tissue volumes are taken on serial sections, the vertically stacked serial images, as well as the serial sections of H&E and CyCIF, all require a pixel-wise alignment such that the stacked section images will form a spatially continuous tissue volume. This process of aligning unaligned tissue section images obtained from different acquisitions is called registration. To register all the H&E together, I used the tissue section from the center of the tissue volume as the baseline target for registration since it will be the most similar to the entire stack of tissue, and this similarity will improve overall registration quality. Registration transforms were calculated between each layer in the stack using Matlab's imregtform function set to affine[83]. The calculated transformations were applied sequentially to all slides, moving from one to the next until all slides were registered to the same coordinates as the central slide (Figure 4.2A), which was chosen as the target because it would maximize similarity to the tissue morphologies at the far ends of the tissue stack.

Although a rougher alignment is acceptable for constructing bulk tissue volumes, it was necessary to have higher level registration between adjacent H&E and CyCIF images when training and testing of deep learning models that translate H&E to CyCIF images. The deep learning models learn based on pixel-level expressions and errors in prediction, so if the H&E and CyCIF images are not aligned well, the models will fail to learn relevant patterns since the errors in prediction will

be in part due to the technical differences of registration. Because of whole slide structural changes that biologically occur in the µm space between sections, it was not possible to register whole slide images adequately without using elastic transformations that operate by deforming the image at the local level as opposed to creating a single set of transform parameters for the whole image, which in this case resulted in imaging artifacts, such as altered stain intensity, blurring, distortion, and warped morphology features, that would skew model training. To get the best registration possible with the least amount of artifacts, I performed fine-tuned CyCIF registration using the same imregtform function in Matlab (set to affine)[83] on smaller cropped ROI images that covered the entire tissue section. Within a single small ROI image, a simple transformation can accurately register the tissue without having to accommodate conflicting transforms from regions located in distant areas of the whole slide. The registration transform for this step was calculated using a manually binarized DAPI image and manually thresholded H&E images as the target, such that the computed transform would best align the nuclei of the two ROIs. No manual registration was performed during this process. After the local ROIs were registered, the paired H&E and CyCIF images were tiled for use in model training.

C. SHIFT models

3D stain propagation models were trained to predict single channel images corresponding to one of the CyCIF stains from input H&E tiles from section 054, e.g. H&E→CD45 or H&E→CD31. These models were built using the architectures previously described[133] and are diagrammed in Table 4.2. Paired H&E and CyCIF

image tiles from section 054 were split into 80% training (8134 tiles) and 20% validation (2034 tiles) sets and each model was trained with a batch size of 4 and learning rate of 0.0002 for 100 epochs. Best models were selected based on the lowest validation loss at each epoch end and were then used for downstream application to held-out H&E WSIs.

Layer	Generator
D1	Conv2d(3, 64, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) LeakyReLU(negative_slope=0.2, inplace=True)
D2	Conv2d(64, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D3	Conv2d(128, 256, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D4	Conv2d(256, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D5	Conv2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D6	Conv2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D7	Conv2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
D8	Conv2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) ReLU(inplace=True)
U1	ConvTranspose2d(512, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U2	ConvTranspose2d(1024, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U3	ConvTranspose2d(1024, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U4	ConvTranspose2d(1024, 512, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U5	ConvTranspose2d(1024, 256, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U6	ConvTranspose2d(512, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U7	ConvTranspose2d(256, 64, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(64, eps=1e-05, momentum=0.1, affine=True) ReLU(inplace=True)
U8	ConvTranspose2d(128, 1, kernel_size=(4,4), stride=(2,2), padding=(1,1)) Tanh()

Layer	Discriminator
1	Conv2d(4, 64, kernel_size=(4,4), stride=(2,2), padding=(1,1)) LeakyReLU(negative_slope=0.2, inplace=True)
2	Conv2d(64, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(128, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
3	Conv2d(128, 256, kernel_size=(4,4), stride=(2,2), padding=(1,1), bias=False) BatchNorm2d(256, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
4	Conv2d(256, 512, kernel_size=(4,4), stride=(1,1), padding=(1,1), bias=False) BatchNorm2d(512, eps=1e-05, momentum=0.1, affine=True) LeakyReLU(negative_slope=0.2, inplace=True)
5	Conv2d(512, 1, kernel_size=(4,4), stride=(1,1), padding=(1,1))

TABLE 4.2: SHIFT model architecture. Layers are represented in PyTorch pseudocode. For the layer column, D and U represent down- and up-sampling layers of the U-Net architecture [103], respectively.

D. Measuring concordance between nuclei overlap in adjacent sections

Because the H&E and CyCIF images are serially sectioned with a thickness of 5 µm, there is a non-negligible difference in tissue architecture imaged on each section. This difference limits the overall evaluation of the models since the predicted CyCIF will line up with the tissue architecture of the H&E image more precisely than with the CyCIF image it is evaluated against. Estimation of the upper bound on SHIFT performance was done by measuring concordance between overlapping nuclei in adjacent sections for locally-registered ROIs from H&E/CyCIF test sections. For H&E ROIs, I deconvolve the hematoxylin stain to extract nuclear content intensity[105], then segment the intensity to derive binary nuclear masks using Cellpose[120]. For CyCIF ROIs, I use Cellpose to segment DAPI intensity to derive binary nuclear masks. The Dice coefficients describing the overlap of nuclear masks from ROIs of adjacent sections were used as compensation factors for evaluating virtual stains. In order to evaluate the overall image reconstruction quality, Dice-compensated structural similarity index measure (SSIM) values were calculated by using scikit-image[139] (with an 11-pixel sliding window). The SSIM between the virtual CyCIF ROI and the real CyCIF ROI were divided by the Dice coefficient of nuclear overlap between the hematoxylin and DAPI nuclear masks from sections 096/097 for that ROI.

4.5.2 Guided region-of-interest selection methods

A. XAE models

In order to select optimal ROIs, it is necessary to quantify feature values for each tissue tile such that I can balance and select relevant features quantitatively. In order to capture histologic features from H&E and expression features from CyCIF in the same latent space, an XAE model was used, which is an encoding and reconstruction model that take two image inputs and co-encodes them into a single descriptor. XAE models were built using Pytorch and are described in Table 4.3. The XAE architecture used here is an adaptation of the UNIT architecture[72] and the imaging-to-omics XAE architecture[110]. XAE models have two input encoders (Figure 4.6), one accepting H&E image tiles (batch size x 3 x 256 x 256), and the other accepting the corresponding paired CyCIF images (batch size x N CyCIF channels x 256 x 256). Both encoders compress their inputs into a shared latent space z. From z, image representations can be upscaled by either H&E or CyCIF decoders. Hence, there are four forward paths through the model: (1) H&E reconstruction: H&E→z→H&E; (2) H&E-to-CyCIF translation: H&E→z→CyCIF; (3) CyCIF reconstruction: CyCIF→z→CyCIF; and (4) CyCIF-to-H&E translation: CyCIF→z→H&E. Models were trained with a batch size of 16 and a learning rate of 0.0001 for 100 epochs. Best models were selected based on the lowest validation loss at each epoch end and were then used for downstream application to held-out H&E WSIs.

Layer	Encoders	Shared?
1	ReflectionPad2d((3, 3, 3, 3)) Conv2d(3, 64, kernel_size=(7,7), stride=(1,1)) InstanceNorm2d(64, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
2	Conv2d(64, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1)) InstanceNorm2d(128, eps=1e-05, momentum=0.1, affine=False) ReLU(inplace=True)	No
3	Conv2d(128, 256, kernel_size=(4,4), stride=(2,2), padding=(1,1)) InstanceNorm2d(256, eps=1e-05, momentum=0.1, affine=False) ReLU(inplace=True)	No
4	ResBlock(N=256 ,K=3, S=1)	No
5	ResBlock(N=256, K=3, S=1)	No
6	ResBlock(N=256, K=3, S=1)	No
z	ResBlock(N=256, K=3, S=1) Reparameterization()	Yes

Layer	Decoders	Shared?
1	ResBlock(N=256, K=3, S=1)	Yes
2	ResBlock(N=256, K=3, S=1)	No
3	ResBlock(N=256, K=3, S=1)	No
4	ResBlock(N=256, K=3, S=1)	No
5	ConvTranspose2d(256, 128, kernel_size=(4,4), stride=(2,2), padding=(1,1)) InstanceNorm2d(128, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
6	ConvTranspose2d(128, 64, kernel_size=(4,4), stride=(2,2), padding=(1,1)) InstanceNorm2d(64, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True) ReflectionPad2d((3, 3, 3, 3))	No
7	Conv2d(64, 3, kernel_size=(7,7), stride=(1,1)) Tanh()	No

Layer	Discriminators	Shared?
1	Conv2d(11, 64, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1)) LeakyReLU(negative_slope=0.2, inplace=True)	No
2	Conv2d(64, 128, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1)) InstanceNorm2d(128, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
3	Conv2d(128, 256, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1)) InstanceNorm2d(256, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
4	Conv2d(256, 512, kernel_size=(4, 4), stride=(2, 2), padding=(1, 1)) InstanceNorm2d(512, eps=1e-05, momentum=0.1, affine=False) LeakyReLU(negative_slope=0.2, inplace=True)	No
5	Conv2d(512, 1, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))	No

ResBlock
ReflectionPad2d((1, 1, 1, 1)) Conv2d(N, N, kernel_size=(K, K), stride=(S, S)) InstanceNorm2d(N, eps=1e-05, momentum=0.1, affine=False) ReLU(inplace=True) ReflectionPad2d((1, 1, 1, 1)) Conv2d(N, N, kernel_size=(K, K), stride=(S, S)) InstanceNorm2d(N, eps=1e-05, momentum=0.1, affine=False)

TABLE 4.3: XAE model architecture. Layers are represented in PyTorch pseudocode.

B. Tile cluster identification

Ultimately, I want to evaluate whether deep learning architectures can recapitulate the biological information of both cell type and pathologist annotated histology. I determined that the best way to validate that the model was capturing biologically relevant information from both modalities (H&E and CyCIF) was to compare the ground truth histologic labels and cell type clusters with the clusters created from the XAE embedding. Histologic labels were created via expert pathologist annotation of relevant tissue types. The pathologist labeled 6 tissue types that were distinct and prominent throughout the sample. The whole slide annotations were then tiled to match the tiles used in the XAE. A tile's ground truth label from the pathologist was determined based on the maximum pixel-wise tissue type within the tile (Figure 4.7). Cell type clusters were determined by k-means clustering the CyCIF expressions then the cell type composition was used to assign a ground truth label to each tile (Figure 4.7). The choice of 7 clusters was determined using the elbow method of the silhouette score. A smaller number of clusters within the elbow was chosen to better match the number of pathologist annotations for consistency in evaluation. 7 clusters were computed for both the standard VAE and the XAE encoding vectors to closely match the number of clusters/tissue types in the ground truth label set, but the higher of the two was used since the XAE encodings had to capture both sets of information.

Several metrics were used to evaluate the models' ability to predict spatially arrange ground truth cluster information throughout the whole tissue (Figure 4.7). Cluster purity was used to evaluate how well the two methodologies were able to

reconstruct the same clusters as ground truth:

$$Purity = \frac{1}{N}\sum_{i=1}^{k}\max(c_i \cap t_j) \tag{4.2}$$

where N is the number of datapoints, k is the number of clusters, c_i is the set of predicted clusters and t_j is the set of ground truth clusters. The sklearn[96] implementation of Normalized Mutual Information (NMI) was used as another metric to evaluate the same question (Equation 4.1).

To evaluate whether the deep learning models capture the same level of feature information as CyCIF staining, I used the pyrcca[12] implementation of canonical correlation on the encoded latent feature space and the paired CyCIF tile-wise expressions. The outputs from this process produced two components shared between the two modalities (Figure 4.7). Quantitatively the correspondence of the two modalities can be measured by the canonical correlation of each component, and qualitatively the correspondence can be observed by the overlap in the scatter plot of the new components.

C. Region-of-Interest selection methodologies

Once it is known that the tiles contain relevant histologic and cell-type information, the question becomes how to select regions from the whole slide that optimize and balance the XAE embedded features with the fewest regions-of-interest (ROIs). Several methods for ROI selection were tested:

a. Random sampling

Random sampling was conducted by randomly drawing a new non-overlapping ROI repeatedly until the desired number of ROIs were obtained. For bulk analysis and comparison, 1000 random combinations of k ROIs were selected where k is the number of ROIs found to be optimal for the other sampling methods.

b. Convex optimization on composition Convex optimization is a method that applies to a subset of non-linear functions, such that the function is twice differentiable at all points within the domain and such that there exists a straight non-intersecting line connecting any two points along the function. Using convex optimization one can identify the minimum number of ROIs to match WSI cellular population by solving:

$$\min_x ||x||_1 \qquad \text{s.t. } b = Ax$$

where b represents the counts of cells across cell types and A is a matrix whose columns each represent an ROI, the row-wise elements of which contain the number of cells for each cell type.

Since I do not have cell type compositions beforehand, I will use the clustering results from the tile-wise XAE encodings of both H&E and CyCIF. The underlying assumption here is that H&E/CyCIF encoding reflects tile-based cell and tissue composition, which I validated in Figure 4.6. For optimization on XAE cluster composition, I solve:

$$\min_x ||x||_1 \qquad \text{s.t. } b = Ax \text{ and s.t. } 0 \leq x \leq 1$$

where $b \in \mathbb{R}^N$, b represents the composition vector of clustered groups within the WSI, each column of $A \in \mathbb{R}^{N \times M}$ represents a possible ROI and each row contains

the percentage of tiles in that ROI for each cluster; N and M represent the number of clusters and the number of possible ROIs in the WSI respectively. Implementation of this function was conducted using the intlinprog function in MATLAB. The function produces a value for each ROI (x_i) that describes its contribution (or importance) to reconstructing the WSI compositions. The threshold of 0.01 was used as the threshold to select all relevant ROIs with a significant contribution. The main issue of this approach is that it often selects homogeneous cell populations as there is no part of the optimization function that encourages diverse ROIs (Figure 4.8).

c. Convex Optimization with Entropy

Since homogeneous and non-diverse ROIs within this dataset were not biologically interesting, it was necessary to encourage the optimization function to select heterogeneous and diverse ROIs that would capture a wide range of cell types and interactions. To optimize both composition and ROI heterogeneity, I additionally take the entropy of the composition vector into account using the convex function:

$$\min_x(||Ax - b||_2 - \lambda Ex) \qquad \text{s.t. } 0 \leq x \leq 1 \text{ and } \sum x = 1$$

where $E \in \mathbb{R}^M$ represents the vector of ROI entropies and λ is a hyperparameter governing the weight of entropy compared to composition matching with regards to the optimization priority. In this experiment, λ was set to 1. Implementation of this function was conducted using CVX[39] in MATLAB. The function produces a value for each ROI (x_i) that describes its contribution (or importance) to reconstructing the WSI compositions. The threshold of 0.01 was applied as cutoff for these contribution scores to select all relevant ROIs. Because entropy was taken

into account in the optimization function, the resulting ROI selections have more heterogeneity while still converging to a good representation (Figure 4.8).

D. Evaluating selected ROIs

The quality of the selected representative ROIs was evaluated based on three metrics: Mean squared error (MSE), Jensen-Shannon Divergence (JSD), and entropy. MSE and JSD between the ROI and WSI compositions were used to evaluate how well the selected ROI compositions match the composition of the WSI. MSE operates on the differences between individual values comparing predicted and ground truth, while JSD operates by comparing the overall distribution of each. Mean squared error was calculated using:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(R_i - W_i)^2 \tag{4.3}$$

where n is the number of predicted clusters, R is the percent composition of each cluster within all selected ROIs combined, and W_i is the percent composition of each cluster within the WSI. JSD was calculated using:

$$JSD = \frac{1}{2}\sum_{i=1}^{n} R_i log_2\left(\frac{R_i}{\frac{1}{2}(R_i + Wi)}\right) + \frac{1}{2}\sum_{i=1}^{n} W_i log_2\left(\frac{W_i}{\frac{1}{2}(R_i + Wi)}\right) \tag{4.4}$$

where n is the number of predicted clusters, R_i is the percent composition of each cluster within all selected ROIs combined, and W_i is the percent composition of each cluster within the WSI. Mean ROI entropy was then used to evaluate the

heterogeneity of the selected ROIs. The mean entropy was calculated using:

$$\text{mean entropy} = \frac{1}{m}\sum_{i=1}^{m}\sum r_i log(r_i) \tag{4.5}$$

where m is the number of selected ROIs and r_i is the percent composition of each cluster within each individual ROI.

4.5.3 Optimized panel selection methods

A. Panel reduction dataset

The dataset used for testing panel reduction methodologies was a breast cancer tissue microarray (TMA) available on synapse from the human tumor atlas network[2]. The TMA dataset is comprised of 88 cores and 6 different cancer subtypes: luminal A, luminal B, luminalB/HER2+, HER2+, triple negative, and invasive lobular carcinoma (ILC). Reference tissue, normal breast, and cell lines are also included. The TMA was imaged using cyclic immunofluorescence with 40 marker channels. The imaging channels were filtered down to 25 channels of interest by removing autofluorescent and duplicate marker channels (see Table 4.1). The stained images were normalized using histogram stretching to the 1^{st} and 99^{th} percentiles, ignoring background area which was thresholded manually. Segmentation was then performed by The HMS Laboratory of Systems Pharmacology using a UNet model[102]. The segmentation resulted in a total of 737,653 single cells images. As discussed in chapter 3, transformational features of single cell images can skew the latent spaces of encoding models like a VAE. For this reason, the single cell images were corrected for rotation by rotating all images such that the major axes of all

cell masks were aligned and were corrected for polar orientation by flipping the images such that the center of staining mass was located in the same quadrant for all cells. By doing this the model can focus on relevant staining information and ignore transformation information that is irrelevant to retaining panel information.

B. Methodologies for selecting the optimal reduced panel

Within a set of markers, intensity information is often correlated when a portion of the proteins of interest operate along the same pathways, are mutually expressed, or are tied to similar phenotypic states. This can be true for markers that localize to different regions of the cell so long as they are correlated in overall expression for different cell states (Figure 4.17). Although some of these correlated stains might be selected for biologically relevant reasons, quantitatively the information from one or more markers can be used to predict the information of another, meaning that they can be reduced. Based on this, there exists an ideal reduced dataset that maximizes the amount of information gained using the fewest markers.

a. Baseline 1-to-1 intensity substitution using reduced panel only In order to create a baseline comparison of reduced panel performance, it is necessary to access the maximum amount of information retained using just a reduced panel without computational inference. Since the metric for evaluation is the correlation between predicted and ground truth marker expression it was necessary to create predictions of withheld markers from the reduced panel without computation. A simple 1-to-1 expression substitution was used for baseline because it is the least computationally intensive method. To do this I computed the correlation for each marker within the full panel set and paired each withheld marker to its highest

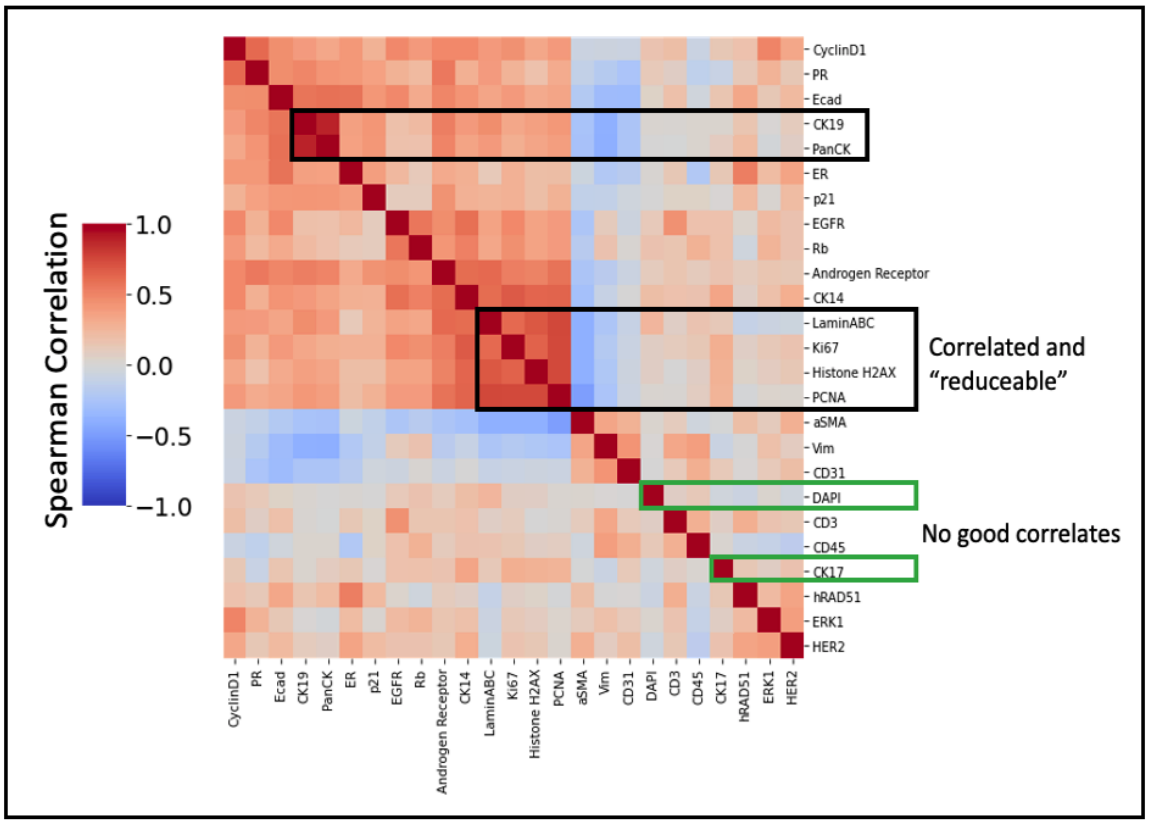


FIGURE 4.17: Heatmap of mean marker intensity correlations in the full TMA panel set, computed across single cell images. Heatmap visualization is clustered using hierarchical clustering of rows and columns. Highly correlated marker clusters show where markers can potentially predict one another and thus can be reduced. Markers with no good correlates will likely need to be included in a reduced panel as there will be no other marker that is predictive of their expression (using intensity information alone). Baseline 1-to-1 substitution will use these correlations to determine marker pairs for intensity substitution. Correlation-based selection will combinatorially create and test all possible panels of size n to determine which reduced panel produces the max correlation to all withheld markers.

correlated partner Figure 4.17 within the randomly selected reduced panel set of a given size. To simulate the predictive inference that would be made by only having the reduced panel set and no reconstruction methods, the predicted intensity for each withheld stain was simply the intensity of its matched partner. This produced fairly low correlations for all panel sizes that only converged to 1 when nearly all

the stains were included in the reduced panel (Figure 4.13).

b. Intensity correlation-based selection In order to determine an optimal reduced panel, the stains that maximized the correlation to all the stains withheld from the panel were chosen. If the correlations between all the stains in the dataset are pre-computed (Figure 4.17), one can quickly perform a combinatorial test of all possible reduced panels with n markers. For every combinatorially created potential panel, the withheld markers are paired with their highest correlated marker within the potential panel, and the max correlations for all withheld markers are averaged to assign a score to that potential panel:

$$PanelScore = \frac{\sum max\left[corr(W_i, R)\right]}{length(W)} \tag{4.6}$$

where R is the potential reduced panel being evaluated, W is the set of withheld markers, and W_i is the intensities of each withheld marker that is being paired. Once I have scored every potential panel of a given size, I can then select the panel that has the greatest score, indicating its predictive capacity toward the withheld markers. Although this method is simplistic, utilizing only mean intensity information, it is quick and is not computationally intensive, making it amenable to rapid panel design and testing. Using this method, the markers are re-selected for each panel size, meaning a specific marker might be included in a panel of one size but not in the next. It is worth noting that for all selection methods, the DAPI marker was a requirement for inclusion in the panel since it is a common marker among currently implemented panels and is a necessary marker for most segmentation pipelines. The panels selected using this method can be seen in Table 4.4.

3-Channel Selection	6-Channel Selection	9-Channel Selection	12-Channel Selection	15-Channel Selection	18-Channel Selection
DAPI	DAPI	DAPI	DAPI	DAPI	DAPI
PCNA	CyclinD1	CyclinD1	hRAD51	CD3	CD3
ECad	VIM	VIM	CyclinD1	ERK-1	ERK-1
	ER	ER	VIM	hRAD51	hRAD51
	PCNA	EGFR	EGFR	CyclinD1	CyclinD1
	PanCK	HER2	HER2	VIM	VIM
		CK17	CD45	aSMA	aSMA
		PCNA	p21	EGFR	ER
		PanCK	CK17	HER2	EGFR
			PCNA	CD45	Rb
			PanCK	p21	HER2
			CD31	CK17	CD45
				PCNA	p21
				PanCK	CK17
				CD31	Androgen Receptor
					PCNA
					PanCK
					CD31

TABLE 4.4: Correlation-based reduced panel set.

c. Sparse subspace clustering-based selection Although the correlation-based method is quick, simple correlation of mean intensities ignores potentially more complex interactions between markers, as two or more markers might be required to predict the expression of a third. Also, combinatorial testing, while quick on small numbers of stains, can become exponentially more burdensome to compute with large panel sets. The next method tested seeks to detect complex interactions across the panel set. To do this deep learning is used to train a coefficient matrix (C) such that the matrix multiplication of C and the single cell marker-wise intensity vector (I) reconstructs I as accurately as possible (Figure 4.18). Here I is an $n \times 1$ matrix where n is the number of markers in the full panel set. During training the diagonal of C is forced to be 0 so that the matrix does not converge to the identity matrix and the off-diagonal coefficients (C_{ij}) give information of the interactions necessary for reconstruction. Training of the C matrix uses the following loss:

$$L_C = \frac{\sum C^2}{n} + ||I - (C \times I)||_2 \tag{4.7}$$

This loss penalizes the size of C such that 1) it remains sparse and only places

weights on a few interactions that contribute the most to reconstruction of the intensity vector and 2) penalizes the accuracy of I reconstruction such that the model learns to compute accurate and relevant interactions (Figure 4.18 right). Looking at the resultant interaction map, one can see that many markers can play a part in predicting the relative intensity of another marker. To determine an optimal panel set from the interaction map, a similar combinatorial method was used, similar to the correlation-based method. Here, however, I was attempting to select a theoretical reduced panel that maximized the interactions to the withheld markers, while minimizing the interactions within the panel:

$$PanelScore = \frac{\sum Int(W_i, R)}{length(W)} - \frac{\sum Int(R_i, R)}{length(R)} \tag{4.8}$$

where R is the potential reduced panel being evaluated, W is the set of withheld markers, W_i is the interactions of each withheld marker to the markers in panel R, and R_i is the interactions of each included marker to the other markers in the reduced panel. Once I have scored every potential panel of a given size, I can then select the panel that has the greatest score, indicating its predictive capacity toward the withheld markers. Just like with correlation-based selection, the reduced set of markers is re-selected for each panel size, meaning a specific marker might be included in a panel of one size but not in the next. The reduced panels selected using this method can be found in Table 4.5

d. Deep learning gradient-based selection While both of the previous methods rely solely on mean intensity information, image data has significantly more information than intensity readouts alone. The localization, texture, and shape of

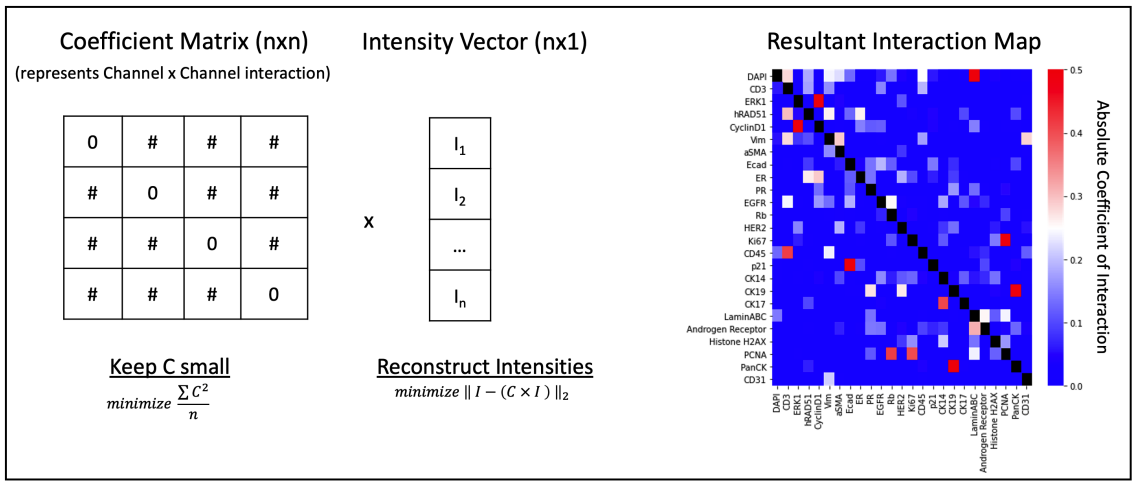


FIGURE 4.18: Diagram demonstrating the trained coefficient matrix and the resultant interaction map used to select a reduced panel. A model is trained to optimize the Coefficient Matrix (C) with a forced zero diagonal, such that it is sparse and when multiplied by the intensity vector of each single cell (I) it can reconstruct I as closely as possible. The resultant interaction map is the trained weights of C, showing the interactions of each marker necessary to adequately reconstruct each other marker in an image. Some makers are capable of being reconstructed from only one other marker, other markers require a more complex combination, and some are not well predicted by any.

cells can also tell us about potential co-staining patterns and interactions. In order to quantify how these features may be important, I use a deep learning method that quantifies the channel-wise importance for reconstructing imaging features across all channels. A similar method to the one described here uses the gradient of the model to determine the channel-wise importance for cell type classification[79]. The key difference of our proposed method is the objective of the model (reconstruction instead of classification) which requires a different architecture. Instead of using a series of ResNet encoders and fully connected classification layers, I use a series of encoders equal to the number of input channels and a decoder for the purpose of reconstructing the combined input channels (Figure 4.19). This forces the gradient at the encoding layer to be greater for channels that play a bigger role

3-Channel Selection	6-Channel Selection	9-Channel Selection	12-Channel Selection	15-Channel Selection	18-Channel Selection
DAPI	DAPI	DAPI	DAPI	DAPI	DAPI
ERK-1	PR	ERK-1	ERK-1	aSMA	CD3
PR	Ki67	PR	aSMA	PR	VIM
	p21	Rb	PR	Rb	aSMA
	CK17	p21	Rb	HER2	Ecad
	CD31	CK17	HER2	Ki67	PR
		Histone H2AX	Ki67	CD45	EGFR
		PanCK	CD45	p21	Rb
		CD31	p21	CK14	Ki67
			CK17	CK17	CD45
			Histone H2AX	LaminABC	p21
			CD31	Androgen Receptor	CK14
				Histone H2AX	CK17
				PCNA	LaminABC
				CD31	Androgen Receptor
					Histone H2AX
					PCNA
					CD31

TABLE 4.5: Subspace-based reduced panel set.

in the reconstruction of the full panel image, including localization, textures, and intensity information. Since the encodings of each channel are kept separate and concatenated, the magnitude of the gradients can be averaged for each channel separately and evaluated for their importance. The reduced panel set is then selected by taking the top n channels ranked by importance, where n is the desired panel size. Because the importance is static and is sequentially added in order or ranking, the selected panels have the advantage of simply being expansions of the smaller panels, potentially making panel design easier. The list of ranked markers can be found in Table 4.6.

Ranked Importance	Marker	Ranked Importance	Marker
1	DAPI	10	ECad
2	CyclinD1	11	VIM
3	CK19	12	ERK-1
4	CK14	13	Rb
5	CK17	14	HER2
6	Ki67	15	CD31
7	EGFR	16	PanCK
8	CD3	17	LaminABC
9	PR	18	Histone H2AX

TABLE 4.6: Gradient-based reduced panel set.

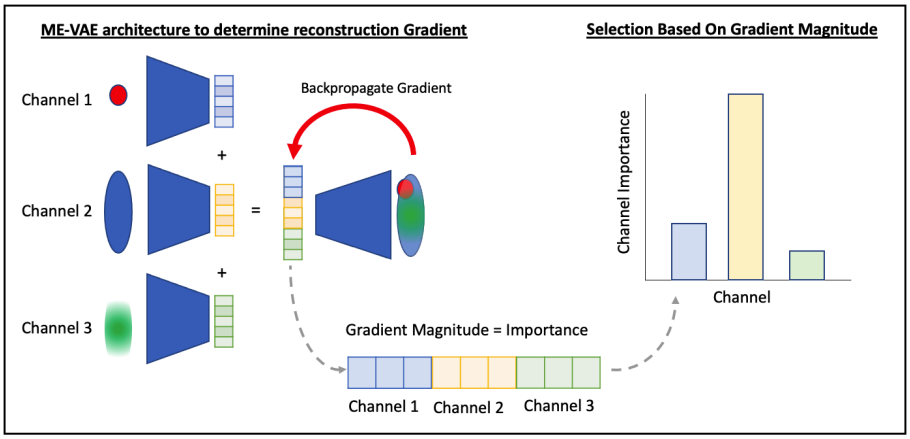


FIGURE 4.19: A multi encoder variational autoencoder architecture is implemented with each channel being used as the input to parallel encoders. The encodings of each channel are concatenated and decoded into a full panel image. The gradients of the model are then backpropagated to the encoding layer. If the magnitude of the gradient is interpreted as importance, the channel gradients can be averaged across the dataset to determine which markers are most important for reconstructing image features within the model.

e. Random selection In order to determine the importance of intelligently selecting panels, I generated random panels for each panel size to be used as a comparison. For ease of testing, I generated a random sequence of markers, and the markers were sequentially added to the panel in that order. The list of markers used for each randomly selected panel can be seen in Table 4.7.

Selection Order	Marker		Selection Order	Marker
1	DAPI		10	hRAD51
2	ERK-1		11	Histone H2AX
3	Ki67		12	PCNA
4	LaminABC		13	CD3
5	CyclinD1		14	ER
6	p21		15	PanCK
7	Androgen Receptor		16	EGFR
8	VIM		17	CK17
9	ECad		18	CK19

TABLE 4.7: Randomly selected reduced panel set.

C. Model architecture for imputing full image and calculating gradient

In order to impute the full panel image from the reduced panel, I trained a multi-encoder variational autoencoder (ME-VAE) where the inputs to each encoder were the channels of the input set and the output was the full panel image. The encodings from each input were concatenated into a single vector before being passed to the decoder. Each encoder and decoder network was 3 layers deep, and each layer used a rectified linear unit activation, except for the output layer which used a sigmoid activation. The concatenated encoding dimension was kept to ~ 128 for all reduced panel sizes. Each input is encoded into its own latent space, equivalent in length to the latent space of all other inputs, meaning that it was not always possible to get a total concatenated latent space of 128 depending on panel size. For this reason a total latent space as close to 128 was chosen for each panel size. This model uses a modified ME-VAE loss since its purpose is to accurately reconstruct an image instead of encode relevant features:

$$L_{ME-VAE} = BCE(x, p(z_{all})) - \frac{1}{n}(\sum_{1}^{n} KL[q_i(z_i|(x_i))||p(z_{all})]) \tag{4.9}$$

where each encoder's (q_i) individual latent space (z_i) is combined in a concatenation layer to create a mutual latent space (z_{all}), x_i represents a different channel of image x, and n represents the number of markers in the reduced panel. Using the described setup, models were trained for 10 epochs and 90% of the dataset.

D. Methods for simulating technical noise

In order to frame the accuracy of predictions in a biological context, I simulated several types of technical noise commonly found in imaging data (blurring, salt/pepper, and variation in segmentation method). To simulate image blurring, I used the scikit-image implementation of gaussian blur [138] with sigma set to 1. To simulate salt/pepper noise, I used the scikit-image implementation of random_noise with the mode set to "s&p" and the amount set to 0.1[138]. To simulate variation in segmentation, I applied erosions and dilations to image masks, eroding half the image and dilating the other half. This created a mask deformed from the original by a few pixels, as can reasonably be expected from different segmentation methods. Images were then re-extracted using the new masks.

E. Metrics for reduced panel evaluation

Three metrics were used in the evaluation of reduced panels. In order to evaluate the mean marker intensity predictions, Spearman correlation was used comparing each marker's mean intensity between ground truth and reconstructed images. The Spearman correlation was computed for each stain individually, and then the average correlation across all stains in the set was reported. In order to evaluate the quality of the reconstructed CyCIF image, I computed the SSIM between each single cell image and its reconstruction. This was done for each channel individually, and then the average across all cells and channels was reported. In order to evaluate the retention of information necessary for downstream phenotyping, NMI was computed for the cluster labels created from the full panel intensities

and for the reconstructed intensities of each selection method. For each selection method, the panel size used for this test was 18. For full panel and all selection methods, k-means was used for clustering with k set to 10 (determined using the elbow method on the full panel dataset). UMAP embeddings, calculated from reconstructed intensities, were also used to visualize the clustered data. Cluster labels were colored such that the reduced panel clusters matched to the cell composition of the full panel clusters.

Chapter 5

Conclusion

A path is made by walking on it.

Readings from Chuang Tzu

5.1 Thesis summary

The biomedical field, more so than most others, must deal with problems concerning complex interconnected systems. Even basic imaging modalities allow researchers and clinicians to capture rich organizational and morphological information of the cancer environment and to make some simple holistic conclusions such as the presence, size, and even stage of tumors. As even more powerful imaging modalities like multiplex imaging are developed, however, the burden placed upon researchers and human annotators increases. Although the information in these new image formats are more highly resolved with increased information density[64], the complexity can be too much for annotators to quickly extract pertinent results without the aid of more sophisticated computational methods, and

the reliance on human annotations and decisions leaves analyses vulnerable to researcher bias[31, 149, 27, 111, 59]. Moreover, these new imaging methods come with many limitations of their own, such as increased imaging time, cost, and adverse effects on the tissue[64, 38, 36, 4, 37]. There is no doubt that the development of multiplex tissue imaging will greatly advance the field of cancer systems biology, but without progress on the computational side as well, the full depth of the imaging data cannot be exploited. Using deep learning we can design models and methods to address many of these problems and guide researchers to novel biological insights. Some of the key contributions made by this dissertation include:

1. the development of a semantic tissue segmentation methodology for labeling hitherto inseparable histologically important features of cancer progression and predicting stain distributions without the effects of uneven staining and biased thresholding (chapter 2).

2. the creation of a novel deep learning architecture to extract biologically relevant features from multiplex images while ignoring previously unavoidable and uninteresting transformational features, resulting in improved performance and analysis (chapter 3).

3. the optimization of the multiplex imaging pipeline through predictive staining, guided region selection, and optimized panel reduction, such that multiplex imaging maximizes the amount of information gained with the least amount of wasted time, effort, and money (chapter 4).

Although the work and methods presented in this dissertation are still experimental, they serve as a proof-of-concept for the advancements that will be developed in

the coming years. Whether in the same form as the methods previously described or in the form of even newer more advanced methods, deep learning will play an important role in multiplex image analysis.

5.2 Significance of the presented work

Although these multiplex imaging modalities are capable of producing unprecedented amounts of information, the actual process of running them is expensive; a new section of tissue can require days to weeks to image, depending on the size of the tissue and the number of stains[64, 131]. If only a small section of the tissue is informative, though, staining the rest of the tissue is a superfluous waste of resources. The same can be said for the marker panels used by researchers: every additional round of staining contributes to negative staining artifacts like autofluorescence[53] and tissue degradation[64], and the amount of information gained from some markers might be negligible when included alongside co-expressed markers. The problem with both of these limitations is that researchers cannot with certainty know which areas of a tissue are biologically interesting and cannot intuitively know which markers are imputable. Furthermore, the increased depth of the data presents increased computational burden. The immediate benefits of addressing these problems are fairly obvious at a glance: decreased imaging time increases the number of tissues that can be analyzed; reduced operational cost enables the funding of more research; and guided sampling to reduce dataset redundancy enables better computation with more accurate and informative results. Additionally there are many more nuanced benefits of this work which are

described below.

Although the primary function of principal investigators is to produce quality research and push biological innovation, most would say that they spend a disproportionate amount time applying for funding, managing budgets, and waiting for datasets to be created. Additionally the quality of their research would be improved if they had more resources. These burdens, however, not only affect the quality of immediate projects, but also affect the long term state of the biomedical research field, as many graduates cite their frustrations with budget and data limitations as key contributors to their decision to opt for alternate career paths outside of academia[73]. Connected to this point, is the desire to feel that one's work is valuable. One of the primary motivators for biomedical students and researchers is the feeling that their work is important and has value to the greater community[121]. Although it is an unavoidable truth the research progresses slowly, there is a difference between rigorous testing and tedious labor, and it is important to prioritize time spent doing research in a way that maximizes intellectual rigor.

Many of these issues regarding time and tedious tasks can be alleviated with novel deep learning methods, but the ability to develop such models is not accessible to small labs or those without deep learning expertise. All of these things contribute to reduced work quality and researcher satisfaction, pushing qualified researchers toward industry[73], where resources are often more readily available. This shift will in the coming years shape the landscape of biomedical advancement. The vast majority of biomedical research in the industrial setting over the past decade has been focused on drug development and testing with less and less research being done of the fundamentals of biological discovery and mechanisms of disease[73].

A balance of these two facets is essential. The fundamentals of discovery are necessary for the development of novel drugs, and the movement of researchers away from academia to better funded industry positions will have a lasting impact on the advancement of the field. This trend cannot be solely attributed to the cost, time, and burdensome limitations of multiplex imaging, but the increased value provided by accessible deep learning will provide researchers with a more amenable environment to perform their experiments and maximize their value.

Although big data is useful because more data inherently has more information and potential findings, it also comes with many downsides as well. In any dataset there are potentially confounding factors and false associations, but studies using big data are more likely to find themselves falling prone to these errors[49] because there is an increased number of variables that need to be scrutinized. This is even more true in imaging data where there is not a universal and standardized set of variables. Imaging features are often difficult to identify, extract, and quantify, meaning that researchers might not even know that there is a confounding factor in the analysis[124]. While conventional analyses are heavily dependent of human-in-the-loop methods (i.e. normalization, thresholding, segmentation, feature selection, noise identification and removal)[123, 80, 22], human-in-the-loop methods are incompatible with big data since the manual operations would need to be performed exponentially many times. This necessitates the implementation of deep learning into general practice so that analysis is not dependent on repetitive researcher input.

Without pre-processing guidance and computational data trimming provided by

deep learning, multiplex imaging could one day face the same fate of genome-wide association studies. Even though the idea of testing every nucleotide in the human genome to identify the source of a disease sounds promising, these studies are held back by the low statistical power of testing hundreds of thousands of variables simultaneously[122]. Studies of this nature that use the standard statistical norms will yield $\sim 25,000$ false-positives on average[49]. To overcome this, genome studies require extremely large sample sizes. We must be even more wary of these errors with multiplex imaging which has the false appearance of being a large diverse dataset comprised of hundreds of thousands of cells. In actuality multiplex imaging studies are fairly small, typically containing only a few patient or disease samples[20]. One reason for this small dataset size is again the prohibitive time and cost of acquiring multiplex images, making it difficult to rapidly deploy on any and all samples of interest. This will result in findings that appear significant but might not generalize well.

By improving upon and deploying the methods discussed in this dissertation, we can:

1. reduce the resources and redundancy of producing multiplex images
2. increase the time spent doing actual research
3. decrease researcher frustration with tedious tasks
4. identify the most relevant and informative data quickly
5. reduce the amount of redundant, confounding, and uninformative information

6. guide researchers toward novel metrics
7. improve the significance of findings
8. enable new analytical and multimodal methods that were hitherto unavailable
9. allow multiple labs and institutions to achieve similar quantifications with less variation
10. improve potential treatments and ultimately better the lives of patients everywhere

5.3 Limitations of the proposed methods

The methods discussed in this dissertation are shown to be effective within the context of their design and within the datasets they were tested on, but their success is still limited to the research setting, and there are a plethora of improvements that can be made. Here I will briefly discuss some of their limitations and potential opportunities for future research.

5.3.1 Limitations of VISTA

The first limitation of the VISTA tool is the inherent scope of the project in which the tool was developed. VISTA's original purpose was for segmenting and quantifying the abundances of normal acinar, acinar-to-ductal metaplasia, and ductal neoplasia. There are, however, many other important histological features within

the pancreas that are important to researchers in the cancer domain, including lymphatic tissue, islets of Langerhans, desmoplastic stroma, as well as fully developed adenocarcinoma. A fully effective implementation of VISTA would see the inclusion of a wider array of pancreatic tissue features. This would allow its deployment to a variety of experiments outside of constrained cancer progression studies.

The second and more unavoidable limitation of the VISTA tool is the inevitable aging of computation methods as the field progresses. Although the architectures and methods used at the time were satisfactory, many new approaches to segmentation, classification, and object detection have been developed that push the state-of-the-art ever forward[7, 43, 156]. A new implementation of VISTA would likely require the utilization of new architectures, while employing some of the discovered training and normalization techniques described here.

5.3.2 Limitations of the ME-VAE

Variational autoencoders (VAEs) are an important part of the computer vision community, as they and their many derivations are the current standard for extracting descriptive features from images that are not easily quantified[58]. VAEs themselves have many limitations that the community has acknowledged, and because the ME-VAE is a VAE derivation, it suffers from many of the same limitations. That being said, the ME-VAE architecture was designed to overcome one key limitation of VAEs (the hypersensitivity of the model to dominant uninformative transformational features in a datase[44]) and was not intended to overcome the other limitations at this time.

One of the limitations of VAEs is that the latent spaces are often entangled, making it difficult to attribute meaning to any singular encoding feature[44, 86]. Even methods that attempt to force the disentangling of the encodings (the β-VAE[44] and the invariant C-VAE[86]) still suffer from the fact that the encodings can be difficult to interpret in a biological context. Additionally, VAEs are sensitive to hyperparameters that effect its ability to converge and learn meaningful representations[44]. The ME-VAE architecture, as described here, suffers from both of these limitations, being unable to disentangle and interpret the latent space. Addressing these limitations, however, was outside of the scope of the project. Furthermore, the ME-VAE method of using multiple encoders can be added onto other VAE derivations, meaning that when a new architecture is developed that can address the above limitations, the multiple encoder aspect can be incorporated to remove transformational noise.

The ME-VAE itself is still limited in its ability to remove uninformative transformational features, though, which was the purpose of its development. This is because the ME-VAE relies on transformed images with respect to specific identified features. This means that the feature of disinterest must be 1) known, 2) quantifiable, and 3) transformable in the image. There are many known transformational features which can easily be controlled, and many new features can be added as they are found at the discretion of the researcher. It will be necessary on the part of the researcher to make sure that each feature controlled for by the ME-VAE does not have biological interest or meaning within the context of their experiment. Future directions of research could attempt to design an architecture that learns to

remove uninteresting features without the need for prior knowledge, but my initial opinions on this prospect are reticent since features that might be biologically interesting in one dataset, such as polar orientation and size, might be noise in another.

5.3.3 Limitations of the various techniques for optimizing multiplex pipelines

A. Limitations of 3-dimensional virtual stain propagation

The SHIFT method[133] of virtual staining has been shown to successfully predict several marker distributions in both 2D sections and 3D volumes. The technique, however, is still only highly accurate for some stains that are abundant throughout a tissue[133]. Future iterations of virtual SHIFT staining can look into methods for accurately predicting sparse information without the need to add significant amounts of information.

The proposed method is also limited because the model must trained with at least one section of paired CyCIF and H&E for each tissue volume. This means that to successfully reduce the amount of staining needed overall, at least one section of staining must be performed. With the potential size of 3D volumes being hundreds of sections, performing only single section of CyCIF imaging is an improvement from having to stain many serial tissue sections, but further studies can be conducted on the generalizability of the approach which would allow the model to virtually stain unseen tissue volumes.

B. Limitations of representative ROI selection

The XAE used for embedding histologic and immunofluorescent features relies on similar training methods as SHIFT[133], and thus shares similar limitations. Despite the multiplex dataset having dozens of stains, only eight abundant stains are used for feature embedding. More cell and tissue types could be captured for representation if additional markers were successfully embedded. Further studies can look into ways to extend this embedding methodology to more stains and underrepresented cell types.

An additional limitation of the method is that the optimization function fails to quantify actual biological interest. This project sought to select representative regions that captured all the whole slide features and conveyed the heterogeneity of the whole tissue section. The optimization function, however, is capable of being modified to the desire of the user, optimizing for entropy, certain tissue types, or any other quantifiable metric. Instead of a limitation, this can be seen as a potential customizable feature of the method, wherein the user specifies what kind of regions they desire. Future studies should look into ways to quantify the biological interest of regions, such that the most important regions are recommended for study and analysis.

This method also requires one section of paired H&E and CyCIF, which means it cannot reduce the amount of staining needed within a single section dataset. This does allow for a reduction in time when working on parallel slides wherein the information from the first stained section can guide the staining and analysis of the subsequent. If future studies show that the model is generalizable, then a model

trained previously can be applied to an entirely unseen dataset without having to stain the tissue beforehand.

C. Limitations of optimal panel selection

Although the panel optimization study was performed on a diverse dataset comprised of 6 different breast cancer subtypes[2], more research is needed into the generalizability of proposed panels to unseen datasets and pathologies. New biological environments and samples might contain information that would be eliminated with the proposed panels if a marker has different patterns of expression between the training and test datasets. All of the proposed methods for panel selection also fail to biologically explain why certain markers might be imputable from other markers in the panel. They can order and prioritize stains that are useful for image reconstruction, but they offer little in the way of overarching biological principles that can be learned from and applied to new datasets. The methods also don't take into consideration other complex aspects of panel design, such as marker interaction, overlapping wavelengths, and redundant markers for quality control, which are important considerations when designing a panel[135, 6]. Finally, a VAE-based architecture was used for image reconstruction in this study, as I determined that the embedding space might be useful for analysis and identification of cell types; however, if the goal is to impute the full panel such that no information is lost, a generative model for image-to-image translation (like SHIFT) might be more appropriate in future studies.

5.4 Combating an idealized vision of the future

The research conducted here shows that deep learning has a place in the multiplex imaging pipeline, and research into its application will likely continue for many years to come. Some might see an unavoidable future in which artificial intelligence replaces even the most complex tasks and puts experts out of a job, but if such a future does come to pass, it is a far way off. The current state of deep learning is just as its name implies, capable of learning deep and hidden patterns from data, but nowhere within the trained models and weighted kernels have we been able to create something remotely resembling intelligence or comprehension. In the foreseeable future, the role of deep learning will serve primarily to reduce the burden of tedious and reproducible tasks such as identifying cells within images and classifying cell types. That being said, an ideal goal for the immediate future is one of symbiosis and growth. Deep learning is capable of extracting information from images that humans cannot, but its black box nature means that trust in its conclusions will always need to be validated. There is typically no justification behind its results other than: it is the answer that optimized the loss function during training. Despite deep learning's large capacity for discovery, such unsupported answers are not compatible with blind application into healthcare. This does not mean, however, that deep learning should be avoided; only that it should be used as a guide, replacing burdensome tasks, reducing bias with reproducible and quantifiable predictions, and making suggestions to the researcher that can be deciphered and learned from.

Maybe one day, deep learning will be a miracle technology that is able to detect,

diagnose, track, and treat cancer all in one, and on that day we will have created something we might be able to call a true artificial intelligence. But that day is still a distant dream, and for now we are confined to using deep learning that is computationally useful but not intelligent. In the meantime, until we have the advanced powers of artificial intelligence at our fingertips, we will work towards a future where deep learning improves the multiplex imaging pipeline, making the groundbreaking new imaging modalities more accessible to researchers and clinicians.

Bibliography

[1] URL: lincsproject.org/LINCS/centers/overview.

[2] URL: https://www.synapse.org/?source=post_page----------#!Synapse:syn22041595/wiki/603095.

[3] Rehan Akbani, Patrick Kwok Ng, Henrica M. Werner, Maria Shahmoradgoli, Fan Zhang, Zhenlin Ju, Wenbin Liu, Ji-Yeon Yang, Kosuke Yoshihara, Jun Li, and et al. "A pan-cancer Proteomic perspective on the cancer genome atlas". In: *Nature Communications* 5.1 (2014). DOI: 10.1038/ncomms4887.

[4] Michael Angelo, Sean C. Bendall, Rachel Finck, Matthew B. Hale, Chuck Hitzman, Alexander D. Borowsky, Richard M. Levenson, John B. Lowe, Scot D. Liu, Shuchun Zhao, and et al. "Multiplexed ion beam imaging (MIBI) of human breast tumors". In: *Nature medicine* 20.4 (2014), 436–442. ISSN: 1078-8956. DOI: 10.1038/nm.3488.

[5] Michael Angelo, Sean C Bendall, Rachel Finck, Matthew B Hale, Chuck Hitzman, Alexander D Borowsky, Richard M Levenson, John B Lowe, Scot D Liu, Shuchun Zhao, and et al. "Multiplexed ion beam imaging of human breast tumors". In: *Nature Medicine* 20.4 (2014), 436–442. DOI: 10.1038/nm.3488.

[6] Thomas Myles Ashhurst, Adrian Lloyd Smith, and Nicholas Jonathan King. "High-dimensional fluorescence cytometry". In: *Current Protocols in Immunology* 119.1 (2017). DOI: 10.1002/cpim.37.

[7] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *CoRR* abs/1511.00561 (2015). arXiv: 1511.00561. URL: http://arxiv.org/abs/1511.00561.

[8] Yue Bai and Leo L. Duan. "Tuning-Free Disentanglement via Projection". In: (2019). arXiv: 1906.11732 [stat.ML].

[9] Jayme Garcia Barbedo. "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for Plant Disease Classification". In: *Computers and Electronics in Agriculture* 153 (2018), 46–53. DOI: 10.1016/j.compag.2018.08.013.

[10] Souptik Barua, Penny Fang, Amrish Sharma, Junya Fujimoto, Ignacio Wistuba, Arvind U.K. Rao, and Steven H. Lin. "Spatial interaction of tumor cells and regulatory T cells correlates with survival in non-small cell lung cancer". In: *Lung Cancer* 117 (2018), pp. 73–79. ISSN: 0169-5002. DOI: https://doi.org/10.1016/j.lungcan.2018.01.022. URL: https://www.sciencedirect.com/science/article/pii/S0169500218302368.

[11] Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. "Generalizing RNA velocity to transient cell states through dynamical modeling". In: *Nature Biotechnology* 38.12 (2020), 1408–1414. DOI: 10.1038/s41587-020-0591-3.

[12] Natalia Y. Bilenko and Jack L. Gallant. "Pyrcca: Regularized Kernel Canonical Correlation Analysis in Python and Its Applications to Neuroimaging". In: *Frontiers in Neuroinformatics* 10 (2016). DOI: 10.3389/fninf.2016.00049.

[13] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[14] Jason D. Buenrostro, Beijing Wu, Howard Y. Chang, and William J. Greenleaf. "ATAC-Seq: A method for assaying chromatin accessibility genome-wide". In: *Current Protocols in Molecular Biology* 109.1 (2015). DOI: 10.1002/0471142727.mb2129s109.

[15] Erik A. Burlingame, Mary McDonnell, Geoffrey F. Schau, Guillaume Thibault, Christian Lanciault, Terry Morgan, Brett E. Johnson, Christopher Corless, Joe W. Gray, Young Hwan Chang, and et al. "Shift: Speedy histological-to-immunofluorescent translation of a tumor signature enabled by Deep Learning". In: *Scientific Reports* 10.1 (2020). DOI: 10.1038/s41598-020-74500-3.

[16] Erik A. Burlingame, Mary McDonnell, Geoffrey F. Schau, Guillaume Thibault, Christian Lanciault, Terry Morgan, Brett E. Johnson, Christopher Corless, Joe W. Gray, Young Hwan Chang, and et al. "Shift: Speedy histological-to-immunofluorescent translation of whole slide images enabled by Deep Learning". In: (2019). DOI: 10.1101/730309.

[17] Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, Thomas J. Fuchs, and et al. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide

images". In: *Nature Medicine* 25.8 (2019), 1301–1309. DOI: 10.1038/s41591-019-0508-1.

[18] Raúl Catena, Alaz Özcan, Laura Kütt, Alex Plüss, IMAXT Consortium, Peter Schraml, Holger Moch, and Bernd Bodenmiller. *Highly multiplexed molecular and cellular mapping of breast cancer tissue in three dimensions using mass tomography*. 2020. DOI: 10.1101/2020.05.24.113571. URL: http://biorxiv.org/lookup/doi/10.1101/2020.05.24.113571.

[19] Jie Chang, Xiaoci Zhang, Jie Chang, Minquan Ye, Daobin Huang, Peipei Wang, and Chuanwen Yao. "Brain Tumor Segmentation Based on 3D Unet with Multi-Class Focal Loss". In: *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 2018, pp. 1–5. DOI: 10.1109/CISP-BMEI.2018.8633056.

[20] Young Hwan Chang, Erik A. Burlingame, Joe W. Gray, and Adam A. Margolin. "Shift: Speedy histopathological-to-immunofluorescent translation of whole slide images using conditional generative adversarial networks". In: *Medical Imaging 2018: Digital Pathology* (2018). DOI: 10.1117/12.2293249.

[21] Young Hwan Chang, Koei Chin, Guillaume Thibault, Jennifer Eng, Erik Burlingame, and Joe W. Gray. "Restore: Robust intensity normalization method for multiplexed imaging". In: *Communications Biology* 3.1 (2020). DOI: 10.1038/s42003-020-0828-1.

[22] Young Hwan Chang, Koei Chin, Guillaume Thibault, Jennifer Eng, **Erik A. Burlingame**, and Joe W. Gray. "RESTORE: Robust intEnSiTy nORmalization mEthod for multiplexed imaging". In: *Communications Biology* 3.11 (2020), 1–9. ISSN: 2399-3642. DOI: 10.1038/s42003-020-0828-1.

[23] Eric M. Christiansen, Samuel J. Yang, D. Michael Ando, Ashkan Javaherian, Gaia Skibinski, Scott Lipnick, Elliot Mount, Alison O'Neil, Kevan Shah, Alicia K. Lee, and et al. "In SILICO LABELING: Predicting fluorescent labels in unlabeled images". In: *Cell* 173.3 (2018). DOI: 10.1016/j.cell.2018.03.040.

[24] Meiou Dai, Amal A Al-Odaini, Nadège Fils-Aimé, Manuel A Villatoro, Jimin Guo, Ani Arakelian, Shafaat A Rabbani, Suhad Ali, and Jean Jacques Lebrun. "Cyclin D1 cooperates with p21 to regulate TGFβ-mediated breast cancer cell migration and tumor local invasion". In: *Breast Cancer Research* 15.3 (2013). DOI: 10.1186/bcr3441.

[25] Gajalakshmi Dakshinamoorthy, Jaskirat Singh, Joseph Kim, Nadya Nikulina, Roya Bashier, Sejal Mistry, Maria E. Gallina, Atri Choksi, Meenu Perera, Ashley Wilson, and et al. "Abstract 490: Highly multiplexed single-cell spatial analysis of tissue specimens using codex". In: *Immunology* (2019). DOI: 10.1158/1538-7445.am2019-490.

[26] Jeyapradha Duraiyan, Rajeshwar Govindarajan, Karunakaran Kaliyappan, and Murugesan Palanisamy. "Applications of immunohistochemistry". In: *Journal of Pharmacy & Bioallied Sciences* 4.Suppl 2 (2012), S307–S309. ISSN: 0976-4879. DOI: 10.4103/0975-7406.100281.

[27] Loes CG van den Einden, Joanne A de Hullu, Leon FAG Massuger, Johanna MM Grefte, Peter Bult, Anne Wiersma, Adriana CH van Engen-van

Grunsven, Bart Sturm, Steven L Bosch, Harry Hollema, and et al. "Interobserver variability and the effect of education in the histopathological diagnosis of differentiated vulvar intraepithelial neoplasia". In: *Modern Pathology* 26.6 (2013), 874–880. DOI: 10.1038/modpathol.2012.235.

[28] Issam El Naqa and Martin J. Murphy. "What is machine learning?" In: *Machine Learning in Radiation Oncology* (2015), 3–11. DOI: 10.1007/978-3-319-18305-3_1.

[29] Henrik Failmezger, Sathya Muralidhar, Antonio Rullan, Carlos E. de Andrea, Erik Sahai, and Yinyin Yuan. "Topological Tumor Graphs: A Graph-Based Spatial Model to Infer Stromal Recruitment for Immunosuppression in Melanoma Histology". In: *Cancer Research* 80.5 (2020), 1199–1209. ISSN: 0008-5472, 1538-7445. DOI: 10.1158/0008-5472.CAN-19-2268.

[30] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, and et al. "3D Slicer as an image computing platform for the Quantitative Imaging Network". In: *Magnetic Resonance Imaging* 30.9 (2012), 1323–1341. DOI: 10.1016/j.mri.2012.05.001.

[31] B Franc. "Interobserver and intraobserver reproducibility in the histopathology of follicular thyroid carcinoma". In: *Human Pathology* 34.11 (2003), 1092–1100. DOI: 10.1016/s0046-8177(03)00403-9.

[32] Alejandro Francisco-Cruz, Edwin Roger Parra, Michael T. Tetzlaff, and Ignacio I. Wistuba. "Multiplex immunofluorescence assays". In: *Biomarkers for*

Immunotherapy of Cancer (2019), 467–495. DOI: 10.1007/978-1-4939-9773-2_22.

[33] Michael Gadermayr, Laxmi Gupta, Barbara M. Klinkhammer, Peter Boor, and Dorit Merhof. "Unsupervisedly Training GANs for Segmenting Digital Pathology with Automatically Generated Annotations". In: *International Conference on Medical Imaging with Deep Learning*. PMLR, 2019, 175–184. URL: http://proceedings.mlr.press/v102/gadermayr19a.html.

[34] Rohan Gala, Agata Budzillo, Fahimeh Baftizadeh, Jeremy Miller, Nathan Gouwens, Anton Arkhipov, Gabe Murphy, Bosiljka Tasic, Hongkui Zeng, Michael Hawrylycz, and et al. "Consistent cross-modal identification of cortical neurons with coupled autoencoders". In: (2020). DOI: 10.1101/2020.06.30.181065.

[35] Thomas A. Geddes, Taiyun Kim, Lihao Nan, James G. Burchfield, Jean Y. Yang, Dacheng Tao, and Pengyi Yang. "Autoencoder-based cluster ensembles for single-cell RNA-seq data analysis". In: *BMC Bioinformatics* 20.S19 (2019). DOI: 10.1186/s12859-019-3179-5.

[36] Michael J. Gerdes, Christopher J. Sevinsky, Anup Sood, Sudeshna Adak, Musodiq O. Bello, Alexander Bordwell, Ali Can, Alex Corwin, Sean Dinn, Robert J. Filkins, and et al. "Highly multiplexed single-cell analysis of formalin-fixed, paraffin-embedded cancer tissue". In: *Proceedings of the National Academy of Sciences* 110.29 (2013), 11982–11987. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1300136110.

[37] Charlotte Giesen, Hao A. O. Wang, Denis Schapiro, Nevena Zivanovic, Andrea Jacobs, Bodo Hattendorf, Peter J. Schüffler, Daniel Grolimund, Joachim

M. Buhmann, Simone Brandt, and et al. "Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry". In: *Nature Methods* 11.44 (2014), 417–422. ISSN: 1548-7105. DOI: 10.1038/nmeth.2869.

[38] Yury Goltsev, Nikolay Samusik, Julia Kennedy-Darling, Salil Bhate, Matthew Hale, Gustavo Vazquez, Sarah Black, and Garry P. Nolan. "Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging". In: *Cell* 174.4 (2018), 968–981.e15. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.07.010.

[39] Michael Grant and Stephen Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 2.1*. http://cvxr.com/cvx. Mar. 2014.

[40] Noah F. Greenwald, Geneva Miller, Erick Moen, Alex Kong, Adam Kagel, Christine Camacho Fullaway, Brianna J. McIntosh, Ke Leow, Morgan Sarah Schwartz, Thomas Dougherty, and et al. "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and Deep Learning". In: (2021). DOI: 10.1101/2021.03.01.431313.

[41] Christopher Heje Grønbech, Maximillian Fornitz Vording, Pascal N Timshel, Casper Kaae Sønderby, Tune H Pers, and Ole Winther. "SCVAE: Variational auto-encoders for single-cell gene expression data". In: *Bioinformatics* 36.16 (2020), 4415–4422. DOI: 10.1093/bioinformatics/btaa293.

[42] Xifeng Guo, En Zhu, Xinwang Liu, and Jianping Yin. "Affine equivariant autoencoder". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence* (2019). DOI: 10.24963/ijcai.2019/335.

[43] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. "Mask R-CNN". In: *CoRR* abs/1703.06870 (2017). arXiv: 1703.06870. URL: http://arxiv.org/abs/1703.06870.

[44] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *ICLR*. 2017.

[45] Sunil R. Hingorani, Emanuel F. Petricoin, Anirban Maitra, Vinodh Rajapakse, Catrina King, Michael A. Jacobetz, Sally Ross, Thomas P. Conrads, Timothy D. Veenstra, Ben A. Hitt, and et al. "Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse". In: *Cancer Cell* 4.6 (2003), 437–450. DOI: 10.1016/s1535-6108(03)00309-x.

[46] Sunil R. Hingorani, Lifu Wang, Asha S. Multani, Chelsea Combs, Therese B. Deramaudt, Ralph H. Hruban, Anil K. Rustgi, Sandy Chang, and David A. Tuveson. "TRP53R172H and Krasg12d cooperate to promote chromosomal instability and widely metastatic pancreatic ductal adenocarcinoma in mice". In: *Cancer Cell* 7.5 (2005), 469–483. DOI: 10.1016/j.ccr.2005.04.023.

[47] Geoffrey E. Hinton. "Learning multiple layers of representation". In: *Trends in Cognitive Sciences* 11.10 (2007), 428–434. DOI: 10.1016/j.tics.2007.09.004.

[48] Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. "Transforming auto-encoders". In: *Lecture Notes in Computer Science* (2011), 44–51. DOI: 10.1007/978-3-642-21735-7_6.

[49] J. H. Holmes, J. Sun, and N. Peek. "Technical challenges for Big Data in biomedicine and Health: Data Sources, infrastructure, and analytics". In: *Yearbook of Medical Informatics* 23.01 (2014), 42–47. DOI: 10.15265/iy-2014-0018.

[50] Shixia Huang and Chad Creighton. "Reverse phase protein arrays in signaling pathways: A Data Integration Perspective". In: *Drug Design, Development and Therapy* (2015), p. 3519. DOI: 10.2147/dddt.s38375.

[51] Wolfgang Karl Härdle and Léopold Simar. "Canonical correlation analysis". In: *Applied Multivariate Statistical Analysis* (2019), 431–442. DOI: 10.1007/978-3-030-26006-4_16.

[52] Umair Javaid, Damien Dasnoy, and John A. Lee. "Multi-organ segmentation of chest CT images in Radiation oncology: Comparison of standard and dilated unet". In: *Advanced Concepts for Intelligent Vision Systems* (2018), 188–199. DOI: 10.1007/978-3-030-01449-0_16.

[53] Caitlin J. Jenvey and Judith R. Stabel. "Autofluorescence and non-specific immunofluorescent labeling in frozen bovine intestinal tissue sections: Solutions for multicolor immunofluorescence experiments". In: *Journal of Histochemistry & Cytochemistry* 65.9 (2017), 531–541. DOI: 10.1369/0022155417724425.

[54] Brett E. Johnson, Allison L. Creason, Jayne M. Stommel, Jamie Keck, Swapnil Parmar, Courtney B. Betts, Aurora Blucher, Christopher Boniface, Elmar Bucher, Erik Burlingame, Koei Chin, Jennifer Eng, Heidi S. Feiler, Annette Kolodzie, Ben Kong, Marilyne Labrie, Patrick Leyshock, Souraya Mitri, Janice Patterson, Jessica L. Riesterer, Shamilene Sivagnanam,

Damir Sudar, Guillaume Thibault, Christina Zheng, Xiaolin Nan, Laura M. Heiser, Paul T. Spellman, George Thomas, Emek Demir, Young Hwan Chang, Lisa M. Coussens, Alexander R. Guimaraes, Christopher Corless, Jeremy Goecks, Raymond Bergan, Zahi Mitri, Gordon B. Mills, and Joe W. Gray. "An Integrated Clinical, Omic, and Image Atlas of an Evolving Metastatic Breast Cancer". In: *bioRxiv* (2020). DOI: 10.1101/2020.12.03.408500. eprint: https://www.biorxiv.org/content/early/2020/12/03/2020.12.03.408500.full.pdf. URL: https://www.biorxiv.org/content/early/2020/12/03/2020.12.03.408500.

[55] Nikolai N. Khodarev, Bernard Roizman, and Ralph R. Weichselbaum. "Molecular pathways: Interferon/STAT1 pathway: Role in the tumor resistance to genotoxic stress and aggressive growth". In: *Clinical Cancer Research* 18.11 (2012), 3015–3021. DOI: 10.1158/1078-0432.ccr-11-3225.

[56] Ashley Kiemen, Alicia M. Braxton, Mia P. Grahn, Kyu Sang Han, Jaanvi Mahesh Babu, Rebecca Reichel, Falone Amoa, Seung-Mo Hong, Toby C. Cornish, Elizabeth D. Thompson, and et al. "In situ characterization of the 3D microanatomy of the pancreas and pancreatic cancer at single cell resolution". In: (2020). DOI: 10.1101/2020.12.08.416909. URL: http://biorxiv.org/lookup/doi/10.1101/2020.12.08.416909.

[57] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *arXiv:1312.6114 [cs, stat]* (2013). arXiv: 1312.6114. URL: http://arxiv.org/abs/1312.6114.

[58] Diederik P Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: (2014). arXiv: 1312.6114 [stat.ML].

[59] KOTARO KITAYA and TADAHIRO YASUO. "Inter-observer and intra-observer variability in immunohistochemical detection of endometrial stromal plasmacytes in chronic endometritis". In: *Experimental and Therapeutic Medicine* 5.2 (2012), 485–488. DOI: 10.3892/etm.2012.824.

[60] Laura Kuett, Raúl Catena, Alaz Özcan, Alex Plüss, H. R. Ali, M. Al Sa'd, S. Alon, S. Aparicio, G. Battistoni, S. Balasubramanian, and et al. "Three-dimensional imaging mass cytometry for highly multiplexed molecular and cellular mapping of tissues and the tumor microenvironment". In: *Nature Cancer* (2021). DOI: 10.1038/s43018-021-00301-w.

[61] Amal Lahiani, Irina Klaman, Nassir Navab, Shadi Albarqouni, and Eldad Klaiman. "Seamless virtual whole slide image synthesis and validation using perceptual embedding consistency". In: *IEEE Journal of Biomedical and Health Informatics* 25.2 (2021), 403–411. DOI: 10.1109/jbhi.2020.2975151.

[62] Leeor Langer, Yoav Binenbaum, Leonid Gugel, Moran Amit, Ziv Gil, and Shai Dekel. "Computer-aided diagnostics in digital pathology: Automated Evaluation of early-phase pancreatic cancer in mice". In: *International Journal of Computer Assisted Radiology and Surgery* 10.7 (2014), 1043–1054. DOI: 10.1007/s11548-014-1122-9.

[63] A. T. Lee, W. Chew, C. P. Wilding, N. Guljar, M. J. Smith, D. C. Strauss, C. Fisher, A. J. Hayes, I. Judson, K. Thway, and et al. "The adequacy of tissue microarrays in the assessment of inter- and intra-tumoural heterogeneity of infiltrating lymphocyte burden in leiomyosarcoma". In: *Scientific Reports* 9.1 (2019). DOI: 10.1038/s41598-019-50888-5.

[64] Jia-Ren Lin, Benjamin Izar, Shu Wang, Clarence Yapp, Shaolin Mei, Parin M Shah, Sandro Santagata, and Peter K Sorger. "Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes". In: *eLife* 7 (2018). Ed. by Arup K Chakraborty, Arjun Raj, Carsten Marr, and Péter Horváth, e31657. ISSN: 2050-084X. DOI: 10.7554/eLife.31657.

[65] Jia-Ren Lin, Benjamin Izar, Shu Wang, Clarence Yapp, Shaolin Mei, Parin M Shah, Sandro Santagata, and Peter K Sorger. "Highly multiplexed immunofluorescence imaging of human tissues and tumors using T-cycif and conventional optical microscopes". In: *eLife* 7 (2018). DOI: 10.7554/elife.31657.

[66] Jia-Ren Lin, Shu Wang, Shannon Coy, Madison Tyler, Clarence Yapp, Yu-An Chen, Cody N. Heiser, Ken S. Lau, Sandro Santagata, and Peter K. Sorger. "Multiplexed 3D atlas of state transitions and immune interactions in colorectal cancer". In: (2021). DOI: 10.1101/2021.03.31.437984. URL: http://biorxiv.org/lookup/doi/10.1101/2021.03.31.437984.

[67] Jia-Ren Lin, Shu Wang, Shannon Coy, Madison Tyler, Clarence Yapp, Yu-An Chen, Cody N. Heiser, Ken S. Lau, Sandro Santagata, Peter K. Sorger, and et al. "Multiplexed 3D Atlas of state transitions and Immune Interactions in colorectal cancer". In: (2021). DOI: 10.1101/2021.03.31.437984.

[68] Jia-Ren Lin, Mohammad Fallahi-Sichani, Jia-Yun Chen, and Peter K. Sorger. "Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging". In: *Current Protocols in Chemical Biology* 8.4 (2016), 251–264. DOI: 10.1002/cpch.14.

[69] Dongju Liu and Jian Yu. "Otsu Method and K-means". In: *2009 Ninth International Conference on Hybrid Intelligent Systems*. Vol. 1. 2009, pp. 344–349. DOI: 10.1109/HIS.2009.74.

[70] Jonathan T. Liu, Adam K. Glaser, Kaustav Bera, Lawrence D. True, Nicholas P. Reder, Kevin W. Eliceiri, and Anant Madabhushi. "Harnessing non-destructive 3D pathology". In: *Nature Biomedical Engineering* 5.3 (2021), 203–218. DOI: 10.1038/s41551-020-00681-x.

[71] Jonathan T. C. Liu, Adam K. Glaser, Kaustav Bera, Lawrence D. True, Nicholas P. Reder, Kevin W. Eliceiri, and Anant Madabhushi. "Harnessing non-destructive 3D pathology". In: *Nature Biomedical Engineering* 5.33 (2021), 203–218. ISSN: 2157-846X. DOI: 10.1038/s41551-020-00681-x.

[72] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. "Unsupervised Image-to-Image Translation Networks". In: (2018). arXiv: 1703.00848 [cs.CV].

[73] Joseph Loscalzo. "The NIH budget and the Future of Biomedical Research". In: *New England Journal of Medicine* 354.16 (2006), 1665–1667. DOI: 10.1056/nejmp068050.

[74] David N. Louis, Michael Feldman, Alexis B. Carter, Anand S. Dighe, John D. Pfeifer, Lynn Bry, Jonas S. Almeida, Joel Saltz, Jonathan Braun, John E. Tomaszewski, and et al. "Computational pathology: A path ahead". In: *Archives of Pathology & Laboratory Medicine* 140.1 (2015), 41–50. DOI: 10.5858/arpa.2015-0093-sa.

[75] Steve Lu, Julie E. Stein, David L. Rimm, Daphne W. Wang, J. Michael Bell, Douglas B. Johnson, Jeffrey A. Sosman, Kurt A. Schalper, Robert A. Anders, Hao Wang, and et al. "Comparison of biomarker modalities for predicting response to PD-1/PD-L1 checkpoint blockade". In: *JAMA Oncology* 5.8 (2019), p. 1195. DOI: 10.1001/jamaoncol.2019.1549.

[76] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. "A method for normalizing histology slides for quantitative analysis". In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* (2009). DOI: 10.1109/isbi.2009.5193250.

[77] Adam L. MacLean, Tian Hong, and Qing Nie. "Exploring intermediate cell states through the lens of single cells". In: *Current Opinion in Systems Biology* 9 (2018), 32–41. DOI: 10.1016/j.coisb.2018.02.009.

[78] Anant Madabhushi and George Lee. "Image analysis and machine learning in digital pathology: Challenges and opportunities". In: *Medical Image Analysis*. 20th anniversary of the Medical Image Analysis journal (MedIA) 33 (2016), 170–175. ISSN: 1361-8415. DOI: 10.1016/j.media.2016.06.037.

[79] Salma Abdel Magid, Won-Dong Jang, Denis Schapiro, Donglai Wei, James Tompkin, Peter K. Sorger, and Hanspeter Pfister. "Channel embedding for informative protein identification from highly multiplexed images". In: (2020). DOI: 10.1101/2020.03.24.004085.

[80] "Mahotas: Open source software for Scriptable Computer Vision". In: *Journal of Open Research Software* 1.1 (2013). DOI: 10.5334/jors.ac.

[81] Andriy Marusyk, Doris P. Tabassum, Michalina Janiszewska, Andrew E. Place, Anne Trinh, Andrii I. Rozhok, Saumyadipta Pyne, Jennifer L. Guerriero, Shaokun Shu, Muhammad Ekram, Alexander Ishkin, Daniel P. Cahill, Yuri Nikolsky, Timothy A. Chan, Mothaffar F. Rimawi, Susan Hilsenbeck, Rachel Schiff, Kent C. Osborne, Antony Letai, and Kornelia Polyak. "Spatial Proximity to Fibroblasts Impacts Molecular Features and Therapeutic Sensitivity of Breast Cancer Cells Influencing Clinical Outcomes". In: *Cancer Research* 76.22 (2016), pp. 6495–6506. ISSN: 0008-5472. DOI: 10.1158/0008-5472.CAN-16-1457. eprint: https://cancerres.aacrjournals.org/content/76/22/6495.full.pdf. URL: https://cancerres.aacrjournals.org/content/76/22/6495.

[82] Raphaël Marée, Loïc Rollus, Benjamin Stévens, Renaud Hoyoux, Gilles Louppe, Rémy Vandaele, Jean-Michel Begon, Philipp Kainz, Pierre Geurts, Louis Wehenkel, and et al. "Collaborative analysis of multi-gigapixel imaging data using Cytomine". In: *Bioinformatics* 32.9 (2016), 1395–1401. DOI: 10.1093/bioinformatics/btw013.

[83] MATLAB. *9.7.0.1190202 (R2019b)*. Natick, Massachusetts: The MathWorks Inc., 2018.

[84] Tadashi Matsuo, Hiroya Fukuhara, and Nobutaka Shimada. "Transform invariant auto-encoder". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2017). DOI: 10.1109/iros.2017.8206047.

[85] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. "UMAP: Uniform manifold approximation and projection". In: *Journal of Open Source Software* 3.29 (2018), p. 861. DOI: 10.21105/joss.00861.

[86] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. "Invariant Representations without Adversarial Training". In: (2019). arXiv: 1805.09458 [cs.LG].

[87] Kaoru Murata, Masakazu Hattori, Norihito Hirai, Yoriko Shinozuka, Hiromi Hirata, Ryoichiro Kageyama, Toshiyuki Sakai, and Nagahiro Minato. "HES1 directly controls cell proliferation through the transcriptional repression of P27 KIP1". In: *Molecular and Cellular Biology* 25.10 (2005), 4262–4271. DOI: 10.1128/mcb.25.10.4262-4271.2005.

[88] Huu-Giao Nguyen, Annika Blank, Heather E. Dawson, Alessandro Lugli, and Inti Zlobec. "Classification of colorectal tissue images from high throughput tissue microarrays by Ensemble Deep Learning Methods". In: *Scientific Reports* 11.1 (2021). DOI: 10.1038/s41598-021-81352-y.

[89] Antonio Nocito, Juha Kononen, Olli-P. Kallioniemi, and Guido Sauter. "Tissue microarrays (tmas) for high-throughput molecular pathology research". In: *International Journal of Cancer* 94.1 (2001), 1–5. DOI: 10.1002/ijc.1385.

[90] Travis Oliphant. *NumPy: A guide to NumPy*. 2006. URL: http://www.numpy.org/.

[91] Chawin Ounkomol, Sharmishtaa Seshamani, Mary M. Maleckar, Forrest Collman, and Gregory R. Johnson. "Label-free prediction of three-dimensional fluorescence images from transmitted-light microscopy". In: *Nature Methods* 15.11 (2018), 917–920. DOI: 10.1038/s41592-018-0111-2.

[92] Fatih Ozsolak and Patrice M. Milos. "RNA sequencing: Advances, challenges and opportunities". In: *Nature Reviews Genetics* 12.2 (2010), 87–98. DOI: 10.1038/nrg2934.

[93] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. "Swapping Autoencoder for Deep Image Manipulation". In: (2020). arXiv: 2007.00653 [cs.CV].

[94] Edwin R. Parra, Naohiro Uraoka, Mei Jiang, Pamela Cook, Don Gibbons, Marie-Andrée Forget, Chantale Bernatchez, Cara Haymaker, Ignacio I. Wistuba, Jaime Rodriguez-Canales, and et al. "Validation of multiplex immunofluorescence panels using multispectral microscopy for immune-profiling of formalin-fixed and paraffin-embedded human tumor tissues". In: *Scientific Reports* 7.1 (2017). DOI: 10.1038/s41598-017-13942-8.

[95] Pushpak Pati, Sonali Andani, Matheus Palhares Viana, Maria Gabrani, Peter Wild, Jan Hendrik Ruschoff, and Matthew Pediaditis. "Deep positive-unlabeled learning for region of interest localization in breast tissue images". In: *Medical Imaging 2018: Digital Pathology* (2018). DOI: 10.1117/12.2293721.

[96] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[97] Mary Linton Peters, Andrew Eckel, Peter P. Mueller, Angela C. Tramontano, Davis T. Weaver, Anna Lietz, Chin Hur, Chung Yin Kong, and Pari

V. Pandharipande. "Progression to pancreatic ductal adenocarcinoma from pancreatic intraepithelial neoplasia: Results of a simulation model". In: *Pancreatology* 18.8 (2018), 928–934. DOI: 10.1016/j.pan.2018.07.009.

[98] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. "Color transfer between images". In: *IEEE Computer Graphics and Applications* 21.4 (2001), 34–41. DOI: 10.1109/38.946629.

[99] Tyler Risom, David R Glass, Candace C Liu, Belén Rivero-Gutiérrez, Alex Baranski, Erin F McCaffrey, Noah F Greenwald, Adam Kagel, Siri H Strand, Sushama Varma, and et al. "Transition to invasive breast cancer is associated with progressive changes in the structure and composition of Tumor Stroma". In: (2021). DOI: 10.1101/2021.01.05.425362.

[100] Yair Rivenson, Tairan Liu, Zhensong Wei, Yibo Zhang, Kevin de Haan, and Aydogan Ozcan. "Phasestain: The digital staining of label-free quantitative phase microscopy images using Deep Learning". In: *Light: Science & Applications* 8.1 (2019). DOI: 10.1038/s41377-019-0129-y.

[101] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E. Zuckerman, Thomas Chong, Anthony E. Sisk, and et al. "Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning". In: *Nature Biomedical Engineering* 3.6 (2019), 466–477. DOI: 10.1038/s41551-019-0362-y.

[102] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Lecture Notes in Computer Science* (2015), 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[103] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *arXiv:1505.04597 [cs]* (2015). arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597.

[104] Orit Rozenblatt-Rosen, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E. Rood, Orr Ashenberg, Ethan Cerami, Robert J. Coffey, Emek Demir, Li Ding, Edward D. Esplin, James M. Ford, Jeremy Goecks, Sharmistha Ghosh, Joe W. Gray, Justin Guinney, Sean E. Hanlon, Shannon K. Hughes, E. Shelley Hwang, Christine A. Iacobuzio-Donahue, Judit Jané-Valbuena, Bruce E. Johnson, Ken S. Lau, Tracy Lively, Sarah A. Mazzilli, Dana Pe'er, Sandro Santagata, Alex K. Shalek, Denis Schapiro, Michael P. Snyder, Peter K. Sorger, Avrum E. Spira, Sudhir Srivastava, Kai Tan, Robert B. West, Elizabeth H. Williams, Denise Aberle, Samuel I. Achilefu, Foluso O. Ademuyiwa, Andrew C. Adey, Rebecca L. Aft, Rachana Agarwal, Ruben A. Aguilar, Fatemeh Alikarami, Viola Allaj, Christopher Amos, Robert A. Anders, Michael R. Angelo, Kristen Anton, Orr Ashenberg, Jon C. Aster, Ozgun Babur, Amir Bahmani, Akshay Balsubramani, David Barrett, Jennifer Beane, Diane E. Bender, Kathrin Bernt, Lynne Berry, Courtney B. Betts, Julie Bletz, Katie Blise, Adrienne Boire, Genevieve Boland, Alexander Borowsky, Kristopher Bosse, Matthew Bott, Ed Boyden, James Brooks, Raphael Bueno, Erik A. Burlingame, Qiuyin Cai, Joshua Campbell, Wagma Caravan, Ethan Cerami, Hassan Chaib, Joseph M. Chan, Young Hwan Chang, Deyali Chatterjee, Ojasvi Chaudhary, Alyce A. Chen, Bob Chen, Changya Chen, Chia-hui Chen, Feng Chen, Yu-An Chen, Milan G. Chheda, Koei Chin,

Roxanne Chiu, Shih-Kai Chu, Rodrigo Chuaqui, Jaeyoung Chun, Luis Cisneros, Robert J. Coffey, Graham A. Colditz, Kristina Cole, Natalie Collins, Kevin Contrepois, Lisa M. Coussens, Allison L. Creason, Daniel Crichton, Christina Curtis, Tanja Davidsen, Sherri R. Davies, Ino de Bruijn, Laura Dellostritto, Angelo De Marzo, Emek Demir, David G. DeNardo, Dinh Diep, Li Ding, Sharon Diskin, Xengie Doan, Julia Drewes, Stephen Dubinett, Michael Dyer, Jacklynn Egger, Jennifer Eng, Barbara Engelhardt, Graham Erwin, Edward D. Esplin, Laura Esserman, Alex Felmeister, Heidi S. Feiler, Ryan C. Fields, Stephen Fisher, Keith Flaherty, Jennifer Flournoy, James M. Ford, Angelo Fortunato, Allison Frangieh, Jennifer L. Frye, Robert S. Fulton, Danielle Galipeau, Siting Gan, Jianjiong Gao, Long Gao, Peng Gao, Vianne R. Gao, Tim Geiger, Ajit George, Gad Getz, Sharmistha Ghosh, Marios Giannakis, David L. Gibbs, William E. Gillanders, Jeremy Goecks, Simon P. Goedegebuure, Alanna Gould, Kate Gowers, Joe W. Gray, William Greenleaf, Jeremy Gresham, Jennifer L. Guerriero, Tuhin K. Guha, Alexander R. Guimaraes, Justin Guinney, David Gutman, Nir Hacohen, Sean Hanlon, Casey R. Hansen, Olivier Harismendy, Kathleen A. Harris, Aaron Hata, Akimasa Hayashi, Cody Heiser, Karla Helvie, John M. Herndon, Gilliam Hirst, Frank Hodi, Travis Hollmann, Aaron Horning, James J. Hsieh, Shannon Hughes, Won Jae Huh, Stephen Hunger, Shelley E. Hwang, Christine A. Iacobuzio-Donahue, Heba Ijaz, Benjamin Izar, Connor A. Jacobson, Samuel Janes, Judit Jané-Valbuena, Reyka G. Jayasinghe, Lihua Jiang, Brett E. Johnson, Bruce Johnson, Tao Ju, Humam Kadara, Klaus Kaestner, Jacob Kagan, Lukas Kalinke, Robert

Keith, Aziz Khan, Warren Kibbe, Albert H. Kim, Erika Kim, Junhyong Kim, Annette Kolodzie, Mateusz Kopytra, Eran Kotler, Robert Krueger, Kostyantyn Krysan, Anshul Kundaje, Uri Ladabaum, Blue B. Lake, Huy Lam, Rozelle Laquindanum, Ken S. Lau, Ashley M. Laughney, Hayan Lee, Marc Lenburg, Carina Leonard, Ignaty Leshchiner, Rochelle Levy, Jerry Li, Christine G. Lian, Kian-Huat Lim, Jia-Ren Lin, Yiyun Lin, Qi Liu, Ruiyang Liu, Tracy Lively, William J.R. Longabaugh, Teri Longacre, Cynthia X. Ma, Mary Catherine Macedonia, Tyler Madison, Christopher A. Maher, Anirban Maitra, Netta Makinen, Danika Makowski, Carlo Maley, Zoltan Maliga, Diego Mallo, John Maris, Nick Markham, Jeffrey Marks, Daniel Martinez, Robert J. Mashl, Ignas Masilionais, Jennifer Mason, Joan Massagué, Pierre Massion, Marissa Mattar, Richard Mazurchuk, Linas Mazutis, Sarah A. Mazzilli, Eliot T. McKinley, Joshua F. McMichael, Daniel Merrick, Matthew Meyerson, Julia R. Miessner, Gordon B. Mills, Meredith Mills, Suman B. Mondal, Motomi Mori, Yuriko Mori, Elizabeth Moses, Yael Mosse, Jeremy L. Muhlich, George F. Murphy, Nicholas E. Navin, Tal Nawy, Michel Nederlof, Reid Ness, Stephanie Nevins, Milen Nikolov, Ajit Johnson Nirmal, Garry Nolan, Edward Novikov, Philipp Oberdoerffer, Brendan O'Connell, Michael Offin, Stephen T. Oh, Anastasiya Olson, Alex Ooms, Miguel Ossandon, Kouros Owzar, Swapnil Parmar, Tasleema Patel, Gary J. Patti, Dana Pe'er, Itsik Pe'er, Tao Peng, Daniel Persson, Marvin Petty, Hanspeter Pfister, Kornelia Polyak, Kamyar Pourfarhangi, Sidharth V. Puram, Qi Qiu, Álvaro Quintanal-Villalonga, Arjun Raj, Marisol Ramirez-Solano, Rumana Rashid, Ashley N. Reeb, Aviv Regev,

Mary Reid, Adam Resnick, Sheila M. Reynolds, Jessica L. Riesterer, Scott Rodig, Joseph T. Roland, Sonia Rosenfield, Asaf Rotem, Sudipta Roy, Orit Rozenblatt-Rosen, Charles M. Rudin, Marc D. Ryser, Sandro Santagata, Maria Santi-Vicini, Kazuhito Sato, Denis Schapiro, Deborah Schrag, Nikolaus Schultz, Cynthia L. Sears, Rosalie C. Sears, Subrata Sen, Triparna Sen, Alex Shalek, Jeff Sheng, Quanhu Sheng, Kooresh I. Shoghi, Martha J. Shrubsole, Yu Shyr, Alexander B. Sibley, Kiara Siex, Alan J. Simmons, Dinah S. Singer, Shamilene Sivagnanam, Michal Slyper, Michael P. Snyder, Artem Sokolov, Sheng-Kwei Song, Peter K. Sorger, Austin Southard-Smith, Avrum Spira, Sudhir Srivastava, Janet Stein, Phillip Storm, Elizabeth Stover, Siri H. Strand, Timothy Su, Damir Sudar, Ryan Sullivan, Lea Surrey, Mario Suvà, Kai Tan, Nadezhda V. Terekhanova, Luke Ternes, Lisa Thammavong, Guillaume Thibault, George V. Thomas, Vésteinn Thorsson, Ellen Todres, Linh Tran, Madison Tyler, Yasin Uzun, Anil Vachani, Eliezer Van Allen, Simon Vandekar, Deborah J. Veis, Sébastien Vigneau, Arastoo Vossough, Angela Waanders, Nikhil Wagle, Liang-Bo Wang, Michael C. Wendl, Robert West, Elizabeth H. Williams, Chi-yun Wu, Hao Wu, Hung-Yi Wu, Matthew A. Wyczalkowski, Yubin Xie, Xiaolu Yang, Clarence Yapp, Wenbao Yu, Yinyin Yuan, Dadong Zhang, Kun Zhang, Mianlei Zhang, Nancy Zhang, Yantian Zhang, Yanyan Zhao, Daniel Cui Zhou, Zilu Zhou, Houxiang Zhu, Qin Zhu, Xiangzhu Zhu, Yuankun Zhu, and Xiaowei Zhuang. “The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution”. In: *Cell* 181.2 (). DOI: 10.1016/j.cell.2020.03.053.

[105] A. C. Ruifrok and D. A. Johnston. "Quantification of histochemical staining by color deconvolution". In: *Analytical and Quantitative Cytology and Histology* 23.4 (2001), 291–299.

[106] Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. "Simple and Effective VAE Training with Calibrated Decoders". In: (2021). arXiv: 2006.13202 [cs.LG].

[107] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. "A comparison of single-cell trajectory inference methods". In: *Nature Biotechnology* 37.5 (2019), 547–554. DOI: 10.1038/s41587-019-0071-9.

[108] Geoffrey Schau, Erik Burlingame, and Young Hwan Chang. "Dissect: Disentangle sharable content for multimodal integration and crosswise-mapping". In: (2020). DOI: 10.1101/2020.09.04.283234.

[109] Geoffrey Schau, Mark A. Dane, Joe W. Gray, Guillaume Thibault, Laura M. Heiser, and Young Hwan Chang. "Variational autoencoding tissue response to microenvironment perturbation". In: *Medical Imaging 2019: Image Processing* (2019). DOI: 10.1117/12.2512660.

[110] Geoffrey F. Schau, Erik A. Burlingame, Guillaume Thibault, Tauangtham Anekpuritanang, Ying Wang, Joe W. Gray, Christopher Corless, and Young H. Chang. "Predicting primary site of secondary liver cancer with a neural estimator of metastatic origin". In: *Journal of Medical Imaging* 7.01 (2020), p. 1. DOI: 10.1117/1.jmi.7.1.012706.

[111] Christoph Schmitz, Hubert Korr, and Helmut Heinsen. "Design-based counting techniques: The real problems". In: *Trends in Neurosciences* 22.8 (1999), p. 345. DOI: 10.1016/s0166-2236(99)01418-6.

[112] Christian M. Schurch, Salil S. Bhate, Graham L. Barlow, Darci J. Phillips, Luca Noti, Inti Zlobec, Pauline Chu, Sarah Black, Janos Demeter, David R. McIlwain, and et al. "Coordinated Cellular Neighborhoods Orchestrate antitumoral immunity at the Colorectal Cancer Invasive Front". In: *Cell* 182.5 (2020). DOI: 10.1016/j.cell.2020.07.005.

[113] Sskipper Seabold and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python". In: *Proceedings of the 9th Python in Science Conference* (2010). DOI: 10.25080/majora-92bf1922-011.

[114] Caglar Senaras, Muhammad Khalid Khan Niazi, Berkman Sahiner, Michael P. Pennell, Gary Tozbikian, Gerard Lozanski, and Metin N. Gurcan. "Optimized generation of high-resolution phantom images using cGAN: Application to quantification of Ki67 breast cancer images". In: *PLOS ONE* 13.5 (2018), e0196846. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0196846.

[115] G Shi, D DiRenzo, C Qu, D Barney, D Miley, and S F Konieczny. *Maintenance of ACINAR cell organization is critical to preventing kras-induced acinar-ductal metaplasia*. 2012. URL: https://www.nature.com/articles/onc2012210.

[116] Nicholas Sofroniew, Kira Evans, Talley Lambert, Juan Nunez-Iglesias, Ahmet Can Solak, Kevin Yamauchi, Genevieve Buckley, Tony Tung, Grzegorz Bokota, Peter Boone, Jeremy Freeman, Hagai Har-Gil, Loic Royer, Shannon Axelrod, jakirkham, Reece Dunham, Pranathi Vemuri, Mars Huang, Hector, Bryant, Ariel Rokem, Justin Kiggins, Hugo van Kemenade, Heath Patterson, Guillaume Gay, Eric Perlman, Davis Bennett,

Christoph Gohlke, Bhavya Chopra, and Alexandre de Siqueira. *napari/napari: 0.3.0*. Version v0.3.0. May 2020. DOI: 10.5281/zenodo.3785908. URL: https://doi.org/10.5281/zenodo.3785908.

[117] Edward C. Stack, Chichung Wang, Kristin A. Roman, and Clifford C. Hoyt. "Multiplexed immunohistochemistry, imaging, and Quantitation: A review, with an assessment of tyramide signal amplification, multispectral imaging and multiplex analysis". In: *Methods* 70.1 (2014), 46–58. DOI: 10.1016/j.ymeth.2014.08.016.

[118] Peter Storz. "Acinar cell plasticity and development of pancreatic ductal adenocarcinoma". In: *Nature Reviews Gastroenterology & Hepatology* 14.5 (2017), 296–304. DOI: 10.1038/nrgastro.2017.12.

[119] Luke Strgar, Eun-Na Kim, Biposa Bose, Luke Ternes, Guillaume Thibault, and Young Hwan Chang. "Evaluation of Nuclei Segmentation in Absence of Ground Truth Labels". In: *Human Tumor Atlas Network: Face2Face* (2021).

[120] Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. "Cellpose: A generalist algorithm for cellular segmentation". In: (2020). DOI: 10.1101/2020.02.02.931238.

[121] "Supplemental material for from bench to bedside: A communal utility value intervention to enhance students' biomedical science motivation". In: *Journal of Educational Psychology* (2015). DOI: 10.1037/edu0000033.supp.

[122] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. "Benefits and limitations of genome-wide association studies". In: *Nature Reviews Genetics* 20.8 (2019), 467–484. DOI: 10.1038/s41576-019-0127-1.

[123] Wenbing Tao, Hai Jin, Yimin Zhang, Liman Liu, and Desheng Wang. "Image Thresholding using graph cuts". In: *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 38.5 (2008), 1181–1195. DOI: 10.1109/tsmca.2008.2001068.

[124] Luke Ternes, Mark Dane, Sean Gross, Marilyne Labrie, Gordon Mills, Joe Gray, Laura Heiser, and Young Hwan Chang. "ME-vae: Multi-encoder variational AutoEncoder for controlling multiple transformational features in single cell image analysis". In: (2021). DOI: 10.1101/2021.04.22.441005.

[125] Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe Gray, John Muschler, and Young Hwan Chang. "Abstract PO-014: VISTA: VIsual Semantic Tissue Analysis for pancreatic disease quantification in murine cohorts". In: *Cancer Research* 81.22 Supplement (2021), PO–014–PO–014. ISSN: 0008-5472. DOI: 10.1158/1538-7445.PANCA21-PO-014. eprint: https://cancerres.aacrjournals.org/content. URL: https://cancerres.aacrjournals.org/content/81/22_Supplement/PO-014.

[126] Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe W. Gray, John Muschler, and Young Hwan Chang. "Vista: Visual semantic tissue analysis for pancreatic disease quantification in murine cohorts". In: *Scientific Reports* 10.1 (2020). DOI: 10.1038/s41598-020-78061-3.

[127] Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe W. Gray, John Muschler, and Young Hwan Chang. "Vista: Visual semantic tissue analysis for pancreatic disease quantification in murine

cohorts". In: *Scientific Reports* 10.1 (2020). DOI: 10.1038/s41598-020-78061-3.

[128] Luke Ternes, Ge Huang, Christian Lanciault, Guillaume Thibault, Rachelle Riggers, Joe W. Gray, John Muschler, and Young Hwan Chang. "VISTA: VIsual Semantic Tissue Analysis for pancreatic disease quantification in murine cohorts". In: *Scientific Reports* 10.11 (2020), p. 20904. ISSN: 2045-2322. DOI: 10.1038/s41598-020-78061-3.

[129] Orit Rozenblatt-Rosen *et al.* "The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution". In: *Cell* 181.2 (2020), pp. 236–249. ISSN: 0092-8674. DOI: https://doi.org/10.1016/j.cell.2020.03.053. URL: https://www.sciencedirect.com/science/article/pii/S0092867420303469.

[130] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, John L Rinn, and et al. "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells". In: *Nature Biotechnology* 32.4 (2014), 381–386. DOI: 10.1038/nbt.2859.

[131] Takahiro Tsujikawa, Sushil Kumar, Rohan N. Borkar, Vahid Azimi, Guillaume Thibault, Young Hwan Chang, Ariel Balter, Rie Kawashima, Gina Choe, David Sauer, and et al. "Quantitative Multiplex Immunohistochemistry Reveals Myeloid-Inflamed Tumor-Immune Complexity Associated with Poor Prognosis". In: *Cell Reports* 19.1 (2017), 203–217. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2017.03.037.

[132] D.A. TUVESON and S.R. HINGORANI. "Ductal pancreatic cancer in humans and mice". In: *Cold Spring Harbor Symposia on Quantitative Biology* 70 (2005), 65–72. DOI: 10.1101/sqb.2005.70.040.

[133] **Erik A. Burlingame**, Mary McDonnell, Geoffrey F. Schau, Guillaume Thibault, Christian Lanciault, Terry Morgan, Brett E. Johnson, Christopher Corless, Joe W. Gray, and Young Hwan Chang. "SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning". In: *Scientific Reports* 10.11 (2020), p. 17507. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74500-3.

[134] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, Nassir Navab, and et al. "Structure-preserving color normalization and sparse stain separation for histological images". In: *IEEE Transactions on Medical Imaging* 35.8 (2016), 1962–1971. DOI: 10.1109/tmi.2016.2529665.

[135] Prashant Vaidyanathan, Evan Appleton, David Tran, Alexander Vahid, George Church, and Douglas Densmore. "Algorithms for the selection of fluorescent reporters". In: *Communications Biology* 4.1 (2021). DOI: 10.1038/s42003-020-01599-5.

[136] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert

Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[137] Juan Carlos Vizcarra, Erik A. Burlingame, Clemens B. Hug, Yury Goltsev, Brian S. White, Darren R. Tyson, and Artem Sokolov. "A community-based approach to image analysis of cells, tissues and tumors". In: *Computerized Medical Imaging and Graphics* 95 (2021), p. 102013. DOI: 10.1016/j.compmedimag.2021.102013.

[138] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. "scikit-image: image processing in Python". In: *PeerJ* 2 (June 2014), e453. ISSN: 2167-8359. DOI: 10.7717/peerj.453. URL: https://doi.org/10.7717/peerj.453.

[139] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, and Tony Yu. "Scikit-image: Image processing in python". In: *PeerJ* 2 (2014). DOI: 10.7717/peerj.453.

[140] Du Wang, Chaochen Gu, Kaijie Wu, and Xinping Guan. "Adversarial neural networks for basal membrane segmentation of microinvasive cervix carcinoma in histopathology images". In: *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*. Vol. 2. 2017, 385–389. DOI: 10.1109/ICMLC.2017.8108952.

[141] Lidong Wang, Huibin Yang, Andrea Zamperone, Daniel Diolaiti, Phillip L. Palmbos, Ethan V. Abel, Vinee Purohit, Igor Dolgalev, Andrew D. Rhim, Mats Ljungman, and et al. "ATDC is required for the initiation of Kras-induced pancreatic tumorigenesis". In: *Genes & Development* 33.11-12 (2019), 641–655. DOI: 10.1101/gad.323303.118.

[142] Tong Wang, Jie Yuan, Jie Zhang, Ran Tian, Wei Ji, Yan Zhou, Yi Yang, Weijie Song, Fei Zhang, Ruifang Niu, and et al. "ANXA2 binds to STAT3 and promotes epithelial to mesenchymal transition in breast cancer cells". In: *Oncotarget* 6.31 (2015), 30975–30992. DOI: 10.18632/oncotarget.5199.

[143] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. "Image quality assessment: From error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4 (2004), 600–612. DOI: 10.1109/tip.2003.819861.

[144] Michael L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

[145] Gregory P. Way, Maria Kost-Alimova, Tsukasa Shibue, William F. Harrington, Stanley Gill, Federica Piccioni, Tim Becker, Hamdah Shafqat-Abbasi, William C. Hahn, Anne E. Carpenter, and et al. "Predicting cell health phenotypes using image-based morphology profiling". In: *Molecular Biology of the Cell* 32.9 (2021), 995–1005. DOI: 10.1091/mbc.e20-12-0784.

[146] Gregory P. Way, Michael Zietz, Vincent Rubinetti, Daniel S. Himmelstein, and Casey S. Greene. "Compressing gene expression data using multiple

latent space dimensionalities learns complementary biological representations". In: *Genome Biology* 21.1 (2020). DOI: 10.1186/s13059-020-02021-3.

[147] Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, and Gene Myers. "Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy". In: *The IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020. DOI: 10.1109/WACV45572.2020.9093435.

[148] Martin C. Whittle and Sunil R. Hingorani. "Understanding disease biology and informing the management of pancreas cancer with preclinical model systems". In: *The Cancer Journal* 23.6 (2017), 326–332. DOI: 10.1097/ppo.0000000000000289.

[149] Suzanne Wilhelmus, H. Terence Cook, Laure-Hélène Noël, Franco Ferrario, Ron Wolterbeek, Jan A. Bruijn, and Ingeborg M. Bajema. "Interobserver agreement on histopathological lesions in class III or IV lupus nephritis". In: *Clinical Journal of the American Society of Nephrology* 10.1 (2014), 47–53. DOI: 10.2215/cjn.03580414.

[150] N. Wissozky. "Ueber das Eosin als Reagens auf Hämoglobin und die Bildung von Blutgefässen und Blutkörperchen bei Säugethier- und Hühnerembryonen". In: *Archiv für mikroskopische Anatomie* 13.1 (1877), 479–496. ISSN: 0176-7364. DOI: 10.1007/BF02933947.

[151] Junyan Wu, Eric Z. Chen, Ruichen Rong, Xiaoxiao Li, Dong Xu, and Hongda Jiang. "Skin Lesion Segmentation with C-UNet". In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 2785–2788. DOI: 10.1109/EMBC.2019.8857773.

[152] Weisi Xie, Adam Glaser, Nicholas Reder, Nadia Postupna, Chenyi Mao, Can Koyuncu, Patrick Leo, Robert Serafin, Hongyi Huang, Anant Madabhushi, and et al. "Abstract PO-017: Annotation-free 3D gland segmentation with generative image-sequence translation for prostate cancer risk assessment". In: *Clinical Cancer Research* 27.5 Supplement (2021), PO–PO–017. ISSN: 1078-0432, 1557-3265. DOI: 10.1158/1557-3265.ADI21-PO-017.

[153] Farhad Ghazvinian Zanjani, Svitlana Zinger, Babak Ehteshami Bejnordi, Jeroen A W M van der Laak, and Peter H. N. de With. "Stain normalization of histopathology images using generative adversarial networks". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. 2018, 573–577. DOI: 10.1109/ISBI.2018.8363641.

[154] Daniel Zhang, Saurabh Mishra, Erik Brynjolfsson, John Etchemendy, Deep Ganguli, Barbara Grosz, Terah Lyons, James Manyika, Juan Carlos Niebles, Michael Sellitto, Yoav Shoham, Jack Clark, and Raymond Perrault. "The AI Index 2021 Annual Report". In: (2021). arXiv: 2103.06312 [cs.AI].

[155] Ziye Zhang, Li Sun, Zhilin Zheng, and Qingli Li. "Disentangling the spatial structure and style in conditional Vae". In: *2020 IEEE International Conference on Image Processing (ICIP)* (2020). DOI: 10.1109/icip40778.2020.9190908.

[156] Zengshun Zhaoa, Yulong Wang, Ke Liu, Haoran Yang, Qian Sun, and Heng Qiao. "Semantic Segmentation by Improved Generative Adversarial Networks". In: *CoRR* abs/2104.09917 (2021). arXiv: 2104.09917. URL: https://arxiv.org/abs/2104.09917.